

Synthese Library 507

Studies in Epistemology, Logic, Methodology,
and Philosophy of Science

Hans J. Briegel
Thomas Müller

Projective Simulation in Action

Quantum-Mechanical Perspectives
on the Problem of Agency

OPEN ACCESS

 Springer

Synthese Library

Studies in Epistemology, Logic, Methodology, and
Philosophy of Science

Volume 507

Editor-in-Chief

Otávio Bueno, Department of Philosophy, University of Miami, Coral Gables, USA

Editorial Board Members

Berit Brogaard, University of Miami, Coral Gables, USA

Steven French, University of Leeds, Leeds, UK

Catarina Dutilh Novaes, VU Amsterdam, Amsterdam, The Netherlands

Darrell P. Rowbottom, Department of Philosophy, Lingnan University, Tuen Mun, Hong Kong

Emma Ruttkamp, Department of Philosophy, University of South Africa, Pretoria, South Africa

Kristie Miller, Department of Philosophy, Centre for Time, University of Sydney, Sydney, Australia

The aim of *Synthese Library* is to provide a forum for the best current work in the methodology and philosophy of science and in epistemology, all broadly understood. A wide variety of different approaches have traditionally been represented in the Library, and every effort is made to maintain this variety, not for its own sake, but because we believe that there are many fruitful and illuminating approaches to the philosophy of science and related disciplines.

Special attention is paid to methodological studies which illustrate the interplay of empirical and philosophical viewpoints and to contributions to the formal (logical, set-theoretical, mathematical, information-theoretical, decision-theoretical, etc.) methodology of empirical sciences. Likewise, the applications of logical methods to epistemology as well as philosophically and methodologically relevant studies in logic are strongly encouraged. The emphasis on logic will be tempered by interest in the psychological, historical, and sociological aspects of science. In addition to monographs *Synthese Library* publishes thematically unified anthologies and edited volumes with a well-defined topical focus inside the aim and scope of the book series. The contributions in the volumes are expected to be focused and structurally organized in accordance with the central theme(s), and should be tied together by an extensive editorial introduction or set of introductions if the volume is divided into parts. An extensive bibliography and index are mandatory.

Hans J. Briegel • Thomas Müller

Projective Simulation in Action

Quantum-Mechanical Perspectives on the
Problem of Agency

Hans J. Briegel 
Institute for Theoretical Physics
University of Innsbruck
Innsbruck, Austria

Thomas Müller 
Department of Philosophy
University of Konstanz
Konstanz, Germany



ISSN 0166-6991

ISSN 2542-8292 (electronic)

Synthese Library

ISBN 978-3-031-98118-0

ISBN 978-3-031-98119-7 (eBook)

<https://doi.org/10.1007/978-3-031-98119-7>

This work was supported by University of Konstanz

© The Editor(s) (if applicable) and The Author(s) 2025. This book is an open access publication.

Open Access This book is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this book are included in the book's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the book's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.

The use of general descriptive names, registered names, trademarks, service marks, etc. in this publication does not imply, even in the absence of a specific statement, that such names are exempt from the relevant protective laws and regulations and therefore free for general use.

The publisher, the authors and the editors are safe to assume that the advice and information in this book are believed to be true and accurate at the date of publication. Neither the publisher nor the authors or the editors give a warranty, expressed or implied, with respect to the material contained herein or for any errors or omissions that may have been made. The publisher remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

If disposing of this product, please recycle the paper.

Preface

Agency is all around us: we are agents, many animals can act, and the future development of Artificial Intelligence may soon lead to the first real artificial agents. But despite its fundamental importance, the notion of agency itself has not been much studied. In philosophy, there has been a long debate about free will, which pertains to the specifically human form of *moral* agency, but the underlying notion of agency has been largely ignored. Even in philosophy of science, scientific agency such as exemplified through experimentation has not been much discussed. And in the sciences themselves, while the perspective of an observer has been acknowledged as important, e.g., for relativity theory and to some extent in quantum mechanics, there is overall hardly any reflection on the scientists' active role in their study of nature.

The situation is changing, albeit slowly. In biology, the usefulness of the notion of agency for understanding biological individuals and their behaviour has been acknowledged (Okasha 2018, 2024), and the notion of agency has consequently been investigated from an evolutionary perspective (Tomasello 2022; Mitchell 2023). In physics, the foundational position of QBism is putting the agent, as user of a theory, centre stage (Fuchs 2010)—even though the QBist notion of agency is still rather thin. Advances in Artificial Intelligence are leading towards autonomous learning agents of various degrees of sophistication (Häse et al. 2019; Rapp et al. 2024). And in philosophy, action theory became a topic of discussion through the study of intentionality, mostly due to Anscombe (1957) and Davidson (1963), and that discussion has meanwhile been expanded from its initial focus on human agency to agency more broadly conceived (Steward 2012; Ferrero 2022).

This book is meant as a contribution to this agential turn in philosophy and in the sciences. It brings together approaches from philosophy, physics, and AI research. These approaches are backed by a concrete and formally defined agency model, Projective Simulation, that is mentioned in the book's title and which was invented by one of the authors (Briegel 2012; Briegel & De las Cuevas 2012). That model draws inspiration from and finds application in quantum physics, highlighting the active role of agents in the scientific practices of that field. At the same time, the model is motivated by a thorough philosophical explication of agency and its place

in nature (Müller & Briegel 2018). That explication addresses the problem of agency also mentioned in our title: While agency is all around us and utterly familiar from our own active lives and from our daily experiences, it is also a philosophical and scientific conundrum that is only beginning to be understood. We discuss Projective Simulation, including a phenomenological reading of that model, with a view to offering steps towards a solution to this problem. On this path, Quantum Mechanics enforces a fundamental reflection on the role of agency in science and provides opportunities for agency modelling.

*

This book is the result of a long and intense interdisciplinary collaboration. It integrates different but interconnected topics: philosophy of science, reduction and emergence, the structure and interpretation of Quantum Mechanics as a theory, QBism as an approach to understanding that theory from a user's point of view, the concrete implementation of Projective Simulation as an agency model, and the role of phenomenology for understanding both human agency and the notion of agency more broadly as it applies to animals and artefacts.

We have discussed agency and its philosophical and physical basis, agency modelling, and applications of Projective Simulation for over a dozen years. In interaction with communities of researchers from the sciences and the humanities and in collaboration with many co-authors, we have embarked on a quest for understanding the problem of agency from many different perspectives. That quest is ongoing, but we have reached a stage at which it seems right to publish our research in the form of a monograph in order to provide a unified account of its motivation, its philosophical and scientific basis, and of our main results. Part of the research that we report on here has been presented in the form of talks or in publications, and we have benefitted greatly from feedback given at colloquia, workshops, and conferences. While we integrate several ideas and results that we have made public before, we have written this book completely from scratch, and we also report on a large body of wholly new research. We present this multidisciplinary research monograph in the hope of enabling a critical discussion of our ideas, welcoming feedback from all the disciplines we address.

Over the course of many years, we have benefitted from grants from a number of institutions whom we wish to thank for their support: the *Templeton World Charity Foundation* (0078/AB46), the *Ministerium für Wissenschaft, Forschung und Kunst Baden-Württemberg* (Az:33-7533.-30-10/41/1), the *University of Innsbruck*, the *University of Konstanz*, the *VolkswagenStiftung* (Az:97721), the *European Research Council* (101055129, PI Briegel), and the *Austrian Science Fund (FWF)* (F7102, F4012, WIT9503323).

It is a pleasure to thank our collaborators on the above-mentioned projects, our co-authors, and the members of our research groups for many helpful discussions, including in many cases feedback on our developing manuscript. Without aiming at an exhaustive list, we would like to thank Gemma De los Coves, Vedran Dunjko, Benjamin Eva, Lukas Fiderer, Nicolai Friis, Stephan Hartmann, Sofiene Jerbi, Hadil Karawani, Marius Krumm, Andrea López Incera, Adi Makmal, Alexey Melnikov,

Gorka Muñoz-Gil, Justus Piater, Hendrik Poulsen Nautrup, Tim Rüz, Katja Ried, Edmund Runggaldier, Daniela Schuster, Daniel Skibra, Isaac Smith, Sahra Styger, Markus Tiersch, Lea Trenkwalder, Maud van Lier, Alexander Vining, and Verena Wagner.

Special thanks are due to Fulvio Flamini and Hendrik Poulsen Nautrup for designing the figures and for helpful feedback on our manuscript, to Sofie Vaas for help with the index, and to Lukas Fiderer and Marius Krumm for detailed comments on previous drafts of several chapters. Thanks also to an anonymous reviewer for helpful comments. Of course, any remaining errors are our own.

We also thank the friendly staff at Springer and our series editor, Otávio Bueno, for their patience and for their support.

H. J. B. would like to thank Karin, Felix and Karl for their love and encouragement.

T. M. thanks J.

Innsbruck, Austria
Konstanz, Germany
March 2024

Hans J. Briegel
Thomas Müller

References

- Anscombe, G. E. M. (1957). *Intention*. Blackwell. 2nd edition 1963, reprint Harvard University Press, Cambridge, MA, 2000.
- Briegel, H. J. (2012). On creative machines and the physical origins of freedom. *Scientific Reports*, 2, 522.
- Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, 2, 400.
- Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60(23), 685–700.
- Ferrero, L. (Ed.). (2022). *The Routledge handbook of philosophy of agency*. Routledge.
- Fuchs, C. A. (2010). QBism, the perimeter of quantum Bayesianism. arXiv 1003.5209.
- Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1, 282–291.
- Mitchell, K. J. (2023). *Free agents. How evolution gave us free will*. Princeton University Press.
- Müller, T., & Briegel, H. J. (2018). A stochastic process model for free agency under indeterminism. *Dialectica*, 72(2), 219–252.
- Okasha, S. (2018). *Agents and goals in evolution*. Oxford University Press.
- Okasha, S. (2024). The concept of agent in biology: Motivations and meanings. *Biological Theory*, 19, 6–10.

- Rapp, J. T., Bremer, B. J., & Romero, P. A. (2024). Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1), 97–107.
- Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
- Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.

Competing Interests The authors have no competing interests to declare that are relevant to the content of this manuscript.

Contents

1	Introduction	1
1.1	The Problem of Agency	1
1.2	Asking the Agency Question with a Straight Face	2
1.3	Modelling Agency as <i>Projective Simulation</i>	3
1.4	Agency, Science, and Quantum Mechanics	6
1.5	Structure of This Book	8
1.5.1	Agency	8
1.5.2	Quantum Mechanics	9
1.5.3	QBism	9
1.5.4	Projective Simulation	10
1.5.5	Phenomenology	11
	References	12
2	Agency	15
2.1	Explicating Agency	16
2.1.1	A Preliminary Account of Agency	17
2.1.2	Topics to Be Addressed in This Chapter	21
2.2	Ontological Levels of Being	22
2.2.1	Ontology and What There Is	22
2.2.2	Reductionism: To Eliminate or not to Eliminate?	25
2.2.3	A Positive Account of Separate Ontological Categories: Ontologically Emergent Individuals	29
2.3	Approaching the Agency Question by Way of Our Practices	42
2.3.1	Agency Talk	45
2.3.2	Agency on the Ladder of Attribution	48
2.3.3	A Feline Example: The Agency of Cats	52
2.4	Attributions of Agency, Their Defeaters, and Agency Scepticism	54
2.4.1	Specific Defeaters of Agency Attribution	55
2.4.2	General Defeaters: Sceptical Doubts	61
2.4.3	Conclusion on Defeaters	65

2.5	The Possibility of Artificial Agents	65
2.5.1	Generic Defeaters for Artificial Agency?	66
2.5.2	Artificial Agency: Attribution, Ignorance, and Emergence	69
2.6	Our Explication of Agency	71
2.6.1	A List of Four Criteria for Agency	71
2.6.2	Agency as Free Agency	75
2.6.3	Agency as Necessarily Social?	78
2.7	Agency and Science	79
2.7.1	A Target for Agency Modelling: Projective Simulation	79
2.7.2	Agency in (Quantum) Science	81
	References	84
3	Quantum Mechanics	91
3.1	Quantum Mechanics: A Brief Historical Introduction	91
3.1.1	Physics Around 1900	92
3.1.2	Atomic Spectra and the Anomalous Zeeman Effect	94
3.1.3	The Stern-Gerlach Experiment	95
3.2	The Mathematical Framework of QM	102
3.2.1	Mathematical Basics: Hilbert Space	103
3.2.2	Interpreting the Formalism: The Basics	108
3.2.3	Quantum Dynamics: Schrödinger vs. Heisenberg Picture	111
3.3	Applying QM: When to Update the Quantum State	116
3.3.1	Picking a Picture	117
3.3.2	Quantum State Update: The Born Rule and When to Use It	118
3.3.3	The Stern-Gerlach Experiment vs. the Mach-Zehnder Interferometer	131
3.3.4	Complementarity and the Heisenberg Uncertainty Relation	149
3.4	The Deterministic Itch	153
3.4.1	How to Deal with Quantum Randomness?	153
3.4.2	Images of a Good Theory	157
3.4.3	Bell's Theorem	163
3.4.4	Quantum Information Theory: Bell-type States as a Resource	174
3.5	Quantum Randomness as a Resource for Sophisticated Agents	186
3.5.1	Randomness for Technological Applications	187
3.5.2	From General Physical to Quantum Random Generators	189
3.5.3	Device-independent Randomness	190
3.5.4	Noise and Quantum Error Correction	191

3.6	Quantum Mechanics and Agency as a Package Deal	205
3.6.1	Lessons from Bell's Theorem	206
3.6.2	QM as a Basis for the Ontological Emergence of Agency	207
	References	210
4	QBism: Integrating Agency and Quantum Mechanics	219
4.1	Integrating Agency and Quantum Physics	220
4.2	Introducing QBism	223
4.2.1	QBism, Quantum Paradoxes, and Common Criticisms	225
4.2.2	Going Beyond QBism	226
4.3	Participatory Realism	231
4.3.1	Realism vs. Anti-Realism in the Philosophy of Science	231
4.3.2	QBism and Participatory Realism	233
4.4	Scientific Agency as Modelling	234
4.4.1	A Case Study of Modelling: Computing Chemical Binding Energies	236
4.4.2	Agents as Modellers, Models as Tools	240
4.4.3	Modelling and QBism	242
4.5	Modelling Scientific Agency	247
4.5.1	Modelling Agency, Modelling Modelling	247
4.5.2	Towards Projective Simulation	248
	References	249
5	Projective Simulation	255
5.1	Informal Description of the Model	256
5.2	Technical Description of the Model	259
5.2.1	Structure of Episodic Memory	261
5.2.2	Dynamics of Deliberation	262
5.2.3	Dynamics of Learning	263
5.2.4	Basic Examples	265
5.3	Contrasting PS with Standard Reinforcement Learning	274
5.4	Applications of PS in Quantum Information Science	281
5.4.1	Quantum Error Correction: In-situ Optimisation of Quantum Codes in Noisy Environments	282
5.4.2	Quantum Communication Through Noisy Channels: Discovery and Variation of Quantum Protocols	287
5.4.3	Measurement-based Quantum Computation: Implicit Model-building of the Dynamics of a Single Qubit	291
5.4.4	Design of Photonic Experiments	297
5.5	Further Developments of PS and Applications in Other Fields of Research	305
5.5.1	Large Percept and Action Spaces	305

5.5.2	The Use of Multi-layered ECMs	306
5.5.3	Quantum Generalisations	308
5.6	Quantum Extensions of PS	308
5.6.1	Quantum Resources for Deliberation and Learning	309
5.6.2	Speed-up of Deliberation via Quantum Walks	310
5.6.3	Quantum Many-body Approach to Deliberation	311
5.6.4	Quantising PS on a Photonic Chip	313
5.7	Interpretability of PS	318
5.7.1	Explanations in Classical PS	318
5.7.2	Human Agents Giving Explanations	320
5.7.3	Transparency and Intransparency in Classical PS	321
5.7.4	Explainability of Quantum PS	322
5.8	Conclusion on PS	324
5.8.1	Modelling Model-building Agents: The Perspective of PS	324
5.8.2	The Status of PS as an Agency Model	326
	References	328
6	Phenomenology	333
6.1	Why Phenomenology?	334
6.2	Being-in-the-world as Acting	336
6.2.1	Being Absorbed: Equipment, Involvement and Reference, Circumspection, and Coping	337
6.2.2	A Graph-theoretical Analogy	338
6.2.3	Equipment and Affordances	340
6.3	Indeterminism and Existential Possibilities	343
6.3.1	Heidegger on Existential Possibilities	345
6.3.2	Thrownness and the Fit of Agential Possibilities with the Environment	347
6.3.3	Projection and Understanding	348
6.4	Learning	350
6.4.1	Learning <i>How-to</i> vs. Learning <i>What-for</i>	352
6.4.2	Indeterminate Equipment and Toys	354
6.5	A Phenomenological Interpretation of Projective Simulation	355
6.5.1	Preliminaries and Basic Framework of PS and Reinforcement Learning	355
6.5.2	Specific Aspects of the PS Framework	357
6.5.3	An Example of Circumspect Coping	359
6.5.4	Developing Circumspection	362
6.5.5	Projection	363
6.5.6	Thrownness	364
6.6	<i>Dasein</i> and Scientific Agency	365
6.6.1	Self-reference and Care	366
6.6.2	Developing Concepts and Models	368
6.6.3	Agency: Both Embedded and Detached	370
	References	373

7 Conclusion 375

References 379

References 381

Name Index 403

Subject Index 415

About the Authors

Hans J. Briegel is Professor of Theoretical Physics at the University of Innsbruck, Austria. His academic career includes postdoctoral positions at Harvard University, Texas A&M University, and the University of Innsbruck. He was one of the founding directors of the Institute for Quantum Optics and Quantum Information of the Austrian Academy of Sciences. He has held visiting positions at the Universities of Bristol and Konstanz and at Tsinghua University. Briegel's main research lies in quantum information science, where his name is associated with concepts such as the one-way quantum computer, the quantum repeater, and quantum reinforcement learning. Beyond his research in quantum physics, he has also been involved in many interdisciplinary projects. He has received research grants and honours including an ERC Advanced Grant (2022), the Wittgenstein Award of the Austrian Science Fund (2023) and the Erwin Schrödinger Award of the Austrian Academy of Sciences (2024). In 2020 he was elected Fellow of the European Laboratory for Learning and Intelligent Systems, and he has been a Corresponding Member of the Austrian Academy of Sciences since 2013.

Thomas Müller is Professor of Theoretical Philosophy at the University of Konstanz, Germany. His academic career includes postdoctoral positions at Freiburg and Bonn and a faculty position at Utrecht, as well as visiting positions at Caltech and at Innsbruck. In his research he specializes in philosophy of agency, philosophy of science, logic, metaphysics, and philosophy of physics. Beyond his philosophical research, he has also worked in interdisciplinary contexts. Müller has been a Principal Investigator of projects funded by the Dutch Research Council, the European Research Council, the German National Science Foundation, and VolkswagenStiftung. He has been an elected member of the German National Academy of Sciences Leopoldina since 2020.

Chapter 1

Introduction



We are agents, and this makes us special. When we act, we make a difference to what goes on in the world, a difference that is specifically up to us. Our causal impact on what happens is different from that of falling rocks, rising tides, or burning forest fires. When we drink coffee, talk to a friend, or take a walk, we act freely, autonomously, and with a purpose of our own. Asked for explanations for why we do what we do, we can provide reasons for our actions. Our causal impact on the world normally makes sense both to us and to others. And as agents, we develop over time: we learn from experience and adapt to varying circumstances in specific ways. We build up our own personal history of actions, experiences, successes, and failures, and this history informs our choices going forward.

We know about these and many more aspects of our agency from the inside as well as from the observation of others. It is, therefore, a basic fact that agency exists in the world, exemplified by us as well as by others with whom we interact. The details of how exactly this works may be hairy, for sure, but the fact that agency exists can hardly be doubted. Or can it?

1.1 The Problem of Agency

The mentioned observations indeed seem to border on platitudes; they characterise agency as firmly rooted in our everyday view of the world. But agency can also appear to be deeply enigmatic. How is it, *really*, that we as autonomous agents can make a difference to what happens in the physical world that we are part of? How do we fit in? Nagel expresses the philosophical challenge of agency as follows:

Everything I do or that anyone else does is part of a larger course of events that no one 'does', but that happens, with or without explanation. Everything I do is part of something I don't do, because I am a part of the world. (Nagel 1986, 114)

When I take a walk, a physical body moves through space, and the motion of that body is very well described by Newtonian laws—my feet push against the ground and the ground pushes back, propelling me forward. When I take a sip of coffee, matter enters my body—matter whose constituents were formed in stars billions of years ago. The fact that matter becomes part of my body for a while has no significance on the grand scale of things, and the chemical interactions that it undergoes in my bowels are the same that happen everywhere. On this view, there is no agent who does anything. The laws of physics or biochemistry apply to everything that happens in the same way. Thus, the notion of agency as something special, as an independent source of meaningful changes in the world, is threatened once one reflects on the fact that agents have to fit into the larger order of nature, in which things simply happen. Zoom out from our everyday perspective, and the special status of our agency is threatened. Stuff happens and that is all. A similar challenge arises when we do not zoom out, but zoom in, considering the many minute entities and processes that constitute our active lives. Paraphrasing Nagel, we could also say that everything I do *has parts* that no one “does”, but that just happen. The energy-generating Krebs cycle in our cells sustains our metabolism whenever we do anything at all, but it is not an active achievement of ours in any way. The same holds for the signalling in our neural pathways, or for the blood flow in our capillaries. The threat, here, is that whatever we do, it will turn out to be wholly constituted by things we don’t do, right down to the level of the interaction of elementary particles. What *really* happens, it seems, is just the aggregate sum of lots of natural happenings. Even a difficult case of morally relevant decision making in the face of uncertainty threatens to dissolve into a host of impersonal, non-agentive physical and biochemical processes.

Viewed in this light, the notion of agency appears to be cornered in an uncomfortable tight spot. The predicament has a name: the disappearance of the agent.¹ How is it that we are agents, given that what we do is but a tiny part of a larger course of affairs that just happens, and is constituted by smaller things that also just happen? How is our agency possible in our natural world?

1.2 Asking the Agency Question with a Straight Face

Like many questions in philosophy, the agency question can be posed in two different ways. The question can be asked “with a straight face”, as part of a quest for understanding—or “with a sceptical sneer”, from a position that is confident that no satisfactory answer can be forthcoming.² Given the sceptical attitude, any attempt at answering the agency question by providing details will only trigger a new round

¹ See Steward (2012, Ch. 3.5). A classic description of the problem (though not exactly in the terms employed here) is given by Velleman (1992).

² We take the framing of these positions from Lear (1984), with thanks to Mulder (2021).

of the question in which the sceptic points to new issues invented on the spot, or to old issues that have not yet been adequately resolved. A full resolution is excluded by design; it has been decided from the start that no answer will be satisfactory.

The alternative is to ask the question “with a straight face”, from a position of really wanting to know how agency is possible. Agency is real, as we all know, being prime examples ourselves. But it is puzzling to understand how it fits into the overall order of things large and small. This certainly calls for a serious investigation. But such an investigation should proceed in an open and cooperative way. This does not mean that the one asking the question may not ask back for more details when an answer has been given, or even reject certain answers as unhelpful or not to the point. The straight-face-attitude does not have to be uncritical. But it treats the question as a serious one that could potentially be answered.

It is to the latter form of the agency question that this book is addressed. We will have some things to say also to those who are sceptical of the very possibility of agency, but we do not assume that we will be able to fully convince them. On the other hand, we do hope to provide something of interest to those who are puzzled by the question of how agency can be a reality, given all that we know about the world and its organisation at large and small scales. In response to this puzzle, we will discuss a number of key aspects of agency from the perspectives of philosophy and of the sciences, and we will propose ways of modelling that phenomenon in a scientifically well-grounded, philosophically motivated manner.

In the natural world that surrounds us, agency comes in many different forms. While we are most intimately familiar with agency from the example of our own human existence, we know that many animals, such as cats and dogs, can also act. For agency modelling, it is useful to focus on agency as a notion that applies to a broad spectrum of cases, not just to human beings. In fact, we are fully aware that our modelling efforts will not amount to a complete elucidation of our own, highly complex human form of agency. There are important aspects of specifically human agency—for example, moral agency—that we will single out as beyond the scope of our investigation, and which we will therefore leave by the wayside. We maintain, however, that our modelling efforts do elucidate key aspects of simpler forms of agency, which are exhibited by us humans in specific circumstances, by many animals and, possibly, also by certain future technological devices.

1.3 Modelling Agency as *Projective Simulation*

In approaching the problem of agency via modelling, we choose a method that is useful both in philosophy and in the sciences. To model a phenomenon is to elucidate its salient features in a way that helps us to understand it. In philosophy, for example, various formal languages have been used to model important aspects of our natural languages, helping us to understand how we reason and how we communicate with each other. Similarly, systems of possible worlds have been used to model a whole range of philosophical notions, from possibility, necessity

and laws of nature to states of belief. In the sciences, modelling is ubiquitous. For example, in biology, the Lotka-Volterra model provides a simple yet helpful picture of predator-prey dynamics, and in physics, the Ising model explains core aspects of ferromagnetism.

Agency modelling comes with challenges that are typical of philosophical modelling and that do not normally come up in scientific modelling: What, exactly, is the target of the modelling effort? What *is* an agent, and how do we know which aspects of agency exactly are the ones that have to be modelled? Those are in themselves philosophical questions. Agency, like ferromagnetism, is a natural phenomenon that is exemplified in many different instances across the world. But unlike ferromagnetism, agency is an extremely complex and philosophically controversial phenomenon, so that it is not straightforward to say what its salient features are. When dealing with ferromagnetism, one can point to a piece of magnetised iron and demand a model for a system behaving *relevantly like that*. The range of phenomena to be covered can fairly easily be agreed upon: for example, the exhibition of a strong magnetic field with certain characteristics, and the existence of a critical Curie temperature T_C above which the spontaneous magnetisation disappears. In cases of doubt, it will also be fairly easy to judge whether a certain property of the system in question is crucial or not. (For example, it turns out that despite the etymology, a ferromagnet does not have to contain iron; it need not even be metallic.) Given such a list of properties and phenomena, one can judge a model by its merits in relation to how well these are captured. This will not single out *the* model of ferromagnetism—there are in fact different ones, with different uses and different degrees of complexity. But the class of relevant examples of the phenomenon—the class of ferromagnets—is well characterised, and there is general agreement on the salient features.

When it comes to agency, however, things are more complicated. One could point to a cat, or to a human being, and demand a model for a system behaving *relevantly like that*: after all, both cats and human beings are agents. But the range of relevant aspects of the phenomenon of agency that an agency model should cover is much harder to determine. What are the core characteristics of agency? As with many other philosophically relevant concepts (think of knowledge, or of the good life), this question is *itself* philosophically loaded. The list of relevant features of agency to be modelled is itself already controversial, independently of the assessment of any concrete agency model. Thus, before the modelling effort starts, or at least in tandem with that effort, we also need to ascertain the relevant aspects of the phenomenon in question. In this book, we devote the whole of Chap. 2 to the task of explicating the notion of agency, so that we can then discuss the question of agency modelling before a rich and hopefully solid background.

A successful agency model has to answer the agency question, the question of how agency is possible in our world, at least in the sense of supporting the reality of agency in our world by showing how agency *could* be possible. A philosophically respectable answer to a “how actually” question does not have to provide a full scientific account—a “how possibly” story is often enough. But an agency model may have a further use, and this is another aspect of agency modelling that separates

it from typical cases of modelling in the sciences. To model is often to represent by formal means, and the properties of the model (which may consist, for example, of a set of equations) are normally completely different from those of the target system (for example, a ferromagnet). The spin interaction equations of the Ising model do not attract iron filings. But of course a model has to capture the salient, often structural properties of the target system that anchor its characteristic behaviour, and so a good model may provide understanding of the sort needed to actually build a system with the characteristics in question. After all, as a famous slogan has it: “If you can’t make one, you don’t know how it works”.³

With this idea in mind, one can see that some models, for example models of computation like the λ -calculus or the universal Turing machine, actually amount to “making one”: arguably, such models of computation are *themselves* also systems that can compute. In this case, modelling seems to amount to building—in stark contrast to models of ferromagnetism or predator-prey dynamics. What about the case of agency: is an agency model itself (possibly, or even necessarily?) a system that can act? This question is complex, like the phenomenon that it is about. But at any rate it makes good sense to ask whether an agency model not just *models* agency, but also *structurally exhibits* agency. If that is so, then an agency model could also provide the basis for building artificial agents. In this way, such a model would not just show how human and animal agency are possible, but would also show how artificial agency can become a reality.

In this book, we discuss *Projective Simulation*, a physics- and philosophy-inspired framework for modelling agency that integrates a basic stochastic dynamics with aspects of episodic memory and reinforcement learning. This modelling effort bridges several disciplines, showing how agency, a philosophically controversial but also well-entrenched concept, can be represented by physically realistic systems with specific, formally specifiable properties, bringing to bear lessons from quantum dynamics and from machine learning.

The name of the model, Projective Simulation (PS), points to two key aspects of agency: its projective structure and the crucial role of simulation. We will unfold both in more detail in the course of this book. After describing the model in formal detail in Chap. 5, we will address projection and simulation further in Chap. 6, where we discuss a phenomenological approach to agency, one of several sources of motivation for the model. Briefly, the model incorporates the idea of projection in the sense that an agent always needs to be ahead of itself. Being thrown into an environment that is fundamentally not of its own making, an agent needs to act based on an implicit understanding of its environment that allows it to project its past experiences into an open future, thereby making use of immediately accessible possibilities. This process of projection in PS incorporates simulation not in the sense of an explicit calculation of possible future states (e.g., with the objective to select the most promising one), but in the more immediate sense of associating

³ This is the title of an influential paper by Dretske (1994).

similarities. In Projective Simulation, an agent makes use of its past learning history to find a next step into the future.

As we will see, modelling agency via Projective Simulation also promises to shed light on the intricate question of the role of agency in science, and especially, in quantum mechanics.

1.4 Agency, Science, and Quantum Mechanics

Understanding agency at the level of a concrete model is a philosophical and a scientific task. Such modelling not only provides an answer to the philosophical question of how agency fits into our world, but also contributes to the understanding of science itself. It is, after all, a fact that our agency has been central for the development of science and of our modern view of the world. Scientific insight is mostly the result of experiments, that is, of people acting and doing things in a highly specific way. The rise of modern science from the seventeenth century on, as highlighted in Bacon's *Novum Organum* (1620), is intimately tied to the experimental method: to the idea that by intervening on what would otherwise be the normal course of nature, we can pose specific questions to nature, in a cooperative struggle for understanding that can yield much deeper insights than pure observation ever could.⁴ Think about the law of freely falling bodies, Galileo's famous result that $s = \frac{1}{2}gt^2$: the distance traveled, s , is proportional (with a factor $\frac{1}{2}g$ that depends on the acceleration due to gravity, g) to the square of the elapsed time t . As Galileo reports, coming up with this law involved many experiments using carefully manufactured bronze balls rolling down an inclined plane covered with polished parchment. It is Galileo's agency and active intervention that gives rise to these highly controlled circumstances in which he can take his measurements. Additionally, a lot of technology, such as a custom-made water clock, is needed for the measurements themselves. Galileo's result is unthinkable as a product of merely observing what is happening in nature anyway: one has no chance to come up with Galileo's law if one merely observes falling leaves and landing birds, falling rain, hail, and snow, people stumbling in the street, or other instances of naturally occurring vertical downward movement. Furthermore, Galileo's account provides an early example of the intimate interrelation between science and technology.

Starting with the recognition of the power of human agency in the study of nature in the seventeenth century, the development of science has led to an immense growth of knowledge about the natural world, to the development of ever more complex technologies, and to a radical transformation of the human way of living and of our planet Earth, both for good and for ill. Quantum mechanics (QM), developed roughly a century ago, illustrates this process in a particularly dramatic way. It is

⁴ See Hacking (1983) for an overview. On the metaphors at play in Bacon's description of this struggle, which is often referred to as "twisting the lion's tail", see Pesic (1999).

well known that QM provides the theoretical foundations for the semiconductor technologies used in computer hardware, LED lights, and solar cells, but also for the nuclear technologies that can be used to manufacture weapons capable of destroying all human life on earth. And the impact of QM is not limited to the practical sphere—the theory also forces us to reconceptualise our understanding of nature and of our place in it. In the course of this book, we will describe the complex interplay of nature and (human) agency that, as we will argue, is baked into the very structure of QM.

Quantum mechanics, as we understand it, fundamentally challenges the tenability of a representational account of nature. Such an account was plausible as long as science could be framed in terms of pure observation. Astronomy and biological taxonomy as well as anatomy were such observational sciences, and astronomy still is to a large extent. Experimental practices challenge that passive account of science as disengaged theory because experimental intervention entangles the experimenter's agency with her natural object of study. The dialectics of that challenge is, however, subtle. Quite often, it is possible to disarm the challenge, as many instances of experimental intervention followed by observation can be redescribed as epistemically distinguished acts of pure observation in which the initial conditions established through the experimenter are marked as freely chosen and thus, as uncorrelated with the past. Indeed, situations in which the absence of correlations with the past is guaranteed can occur naturally, without experimental intervention—witness so-called natural experiments, e.g., in epidemiology or in economy (DiNardo 2016).

Many of the peculiar roles that experiments have played in the study of nature can still be explained even if one understands experiment merely as observation under favourable circumstances. One such role is that experimental intervention allows one to distinguish causation from mere correlation: If one experimentally intervenes on the cause, the effect changes as a consequence, but not so the other way round. The mere observation of cause and effect does not match the epistemic power of experiment in that case. But sometimes cause and effect can still be distinguished in terms of observation—for example, if there is additional data on intermediate variables, or a guarantee of the absence of certain correlations (Pearl 2002; Pearl & Mackenzie 2018).

A non-agentive reading of experiments in terms of mere observation is thus not ruled out on general terms. But as we will argue, quantum experiments are special in that they do not allow for such a reconceptualisation in terms of passive observation. More precisely, we will show that such a reconceptualisation comes at the price of general agency scepticism, which then undermines experimental science wholesale. The viable alternative is to allow for a fundamental role of agency in experimental science: In experiments, agents really choose which questions to pose to nature. In QM, many such questions (mathematically represented as so-called observables) are complementary, so that the respective answers cannot be integrated into a single unified representational view of nature “out there”. The experimenter is *actively* involved in experiments in a fundamental and ineliminable way.

This fact underscores the usefulness of integrating a discussion of agency with a discussion of quantum mechanics. As we will argue in the course of this book, this integration in fact has implications for theory and for practice. On the theory side, QM highlights the possibility of agency in nature and the role of agency in science. And on the practical side, this integration may also provide the basis for the first implementations of artificial agency in a science context.

1.5 Structure of This Book

The book contains this Introduction, five main Chapters on Agency, Quantum Mechanics, QBism, Projective Simulation, and Phenomenology, and a final Chapter drawing conclusions. Here is a brief overview of the contents of the main Chapters.

1.5.1 Agency

In the following Chap. 2 we address the philosophical problem of agency as sketched at the beginning of this introduction. Given our chosen method of agency modelling, it is of fundamental importance to provide an explication of the notion of agency that defines a useful target for our modelling efforts. We start with an initial explication as offered in the groundbreaking and influential study *A Metaphysics for Freedom* by Steward (2012), which centres on animal agency and its basis in what our world is like.

We agree with Steward’s metaphysical outlook and discuss the agency question not as one of epistemology, of characterising a certain stance on natural phenomena, but as a question of ontology, of what there is. Generally put, the fundamental ontological question is this: How does nature, what there is, leave room for real higher-level ontological categories? Agency is just one such higher category. Life is another important one—but both have a somewhat contentious status in philosophy. Using the hopefully uncontroversial example of the sun—a star—as a real higher-level entity, we discuss the interdependence of levels of organisation and argue for a specific notion of ontologically emergent individuals. As we argue, real higher-level entities must be based on lower-level open possibilities. Having established that general basis, we then work our way towards a refined explication of agency, taking inspiration from our practices of agency attribution and tackling the possibility of artificial agents along the way. We characterise agency as free agency that requires open possibilities for the flexible configuration of the agent’s body. In our resulting explication, we also stress one of the most neglected, but crucially important aspects of agency: agents have to be able to learn and to adapt over time. To model agency is, therefore, to model a specific kind of learning.

1.5.2 *Quantum Mechanics*

In Chap. 3, we address quantum mechanics, which is our most successful and versatile account of the basic constituents of matter, of their interaction and of their compounds. We introduce the theory via a sketch of its history and proceed to lay out its mathematical structure in some detail. This provides the background for addressing the important and controversial question of how that structure is to be interpreted. We do not delve into the current foundational controversy in detail, but we briefly contextualise our own position, which is linked to our account of ontological emergence as detailed in Chap. 2, and which is mathematically best expressed in the Heisenberg picture of quantum dynamics. Indeterminism and the fundamental role of irreversible processes stand out as characteristics of our position.

We discuss how that position is challenged by a widespread conviction of determinism, which we see rooted in images of a good theory that developed mainly in the context of (celestial) mechanics. Against this conviction, we argue that indeterminism in the form of quantum randomness has turned from a philosophical riddle to a technological resource that is increasingly being employed by our current technoscientific society. This claim is supported by discussions of several aspects of quantum information science, including quantum error correction, which will be a crucial ingredient in any technologically viable scheme for quantum computation.

Towards the end of the chapter we argue that agency and quantum mechanics come as a package deal, in two ways. For one, quantum mechanics describes the world as offering open possibilities for the future at a fundamental physical level. This shows that our world provides the basic indeterminism that is needed for real higher-order organisation, including the organisation of agents. Of course, this does not in any way show that the *actual* existence of agents in our world is guaranteed. But it does show that general agency scepticism has no basis in physics. Second, we will show that quantum mechanics is best understood as a theory that describes our agential interactions with nature, and not as a theory that represents nature from the point of view of an inert, detached observer. In our view, the mathematical structure of QM, which is uncontested across the various interpretational divides, clearly points in that direction, and the enigmatic aspects of the dynamics of quantum systems are mitigated or even removed once one takes an agency approach to understanding the theory.

1.5.3 *QBism*

In our approach to quantum mechanics, we are fairly closely aligned with one of the main current positions in the quantum foundations debate, QBism. The QBist position takes quantum mechanics, and scientific theories more generally, as

describing the interaction of an agent with the world: A theory mainly offers a way for the agent to manage her expectations about the outcomes of her interactions, not a picture of what nature is really like. In Chap. 4 we introduce QBism briefly and situate the framework within a range of options for integrating agency and quantum mechanics. We agree with QBism on many fundamental issues, but we do not subscribe to QBism wholesale. We agree with QBism’s pragmatic outlook that takes scientific theories to be tools for agents to use, and with the QBist goal of developing a participatory realism. But we are critical of a number of issues that result from the particular idealisations that QBism in its present form subscribes to. Most importantly, we are critical of the structural choice not to resolve the agent nor the world in any detail. And in terms of basic assumptions, we believe that the epistemological foundations that QBism builds on are too narrow to anchor an adequate view of scientific agency.

In our view, the type of learning that characterises scientific agency is best described in a modelling framework. We consider models to be tools that agents can use and develop in dealing with the world. Guided by concrete examples, we describe a scientific modelling framework that allows for loops of three different degrees of difficulty and complexity. These range from quantum state updating, which neatly fits the QBist approach of working within a fixed probability space, via a broader class of updates including so-called Hamiltonian learning, up to the most difficult cases in which true scientific creativity is at play. All three levels of difficulty need to be acknowledged for a full picture of real scientific agency.

At the end of the Chapter, we describe the project of modelling scientific agency itself, which leads to the topics of Chaps. 5 and 6.

1.5.4 Projective Simulation

Our own approach to agency modelling is spelled out in detail in Chap. 5, where we introduce Projective Simulation (PS), a model for agency that incorporates aspects of reinforcement learning and episodic memory, an indeterministic basic dynamics inspired by physical hopping processes studied in quantum optics, and an overall orientation towards agency as a continuous process of interaction with the environment as described by phenomenology. We explain the formal structure of PS in detail and compare it to the more standard paradigm of reinforcement learning. PS stands out mostly due to its specific memory structure (Episodic and Compositional Memory, ECM) and its openness, which makes it apt as an agency model rather than just a learning model: PS does not start from a fixed environment model to be “solved”. The focus is, rather, on the internal dynamics of deliberation, which is resolved as a stochastic process within the ECM, as well as on learning and adaptation over a long time, allowing for significant and even disruptive changes in the environment to be incorporated. Unlike standard reinforcement learning, PS is thus not oriented towards discovering “the optimal policy”. There simply may be

no such thing, and what is best for an agent may vary greatly over time and may crucially depend on the agent's (learning) history.

After discussing the model, we showcase several applications of PS, prominently in basic research in QM and in quantum information. This strengthens our claim about the intimate connection between agency and quantum physics that we already defended in Chaps. 3 and 4. We also consider quantum versions of PS, which promise to show quantum advantages in machine learning while retaining a maximum of transparency. These applications motivate a separate discussion of transparency and traceability in the context of explainable (quantum) AI.

1.5.5 Phenomenology

In our final Chap. 6 we discuss a phenomenological approach to agency based on Heidegger's analysis of *Dasein*, the human way of being-in-the-world. Inspired by the fundamental AI critique of Dreyfus (1992) and by Wheeler (2005), we read Heidegger's *Being and Time* (1927) as offering a description of the human way of being-in-the-world as *acting* in the world. Furthermore, we believe that the analysis is not strictly limited to the specifically *human* kind of agency, but—suitably transposed—also spells out general structures (*existentiales*) of agency of *all* possible forms. The most basic such structure is circumspect coping with the environment, through which an agent interacts with an environment understood as offering sensible possibilities in the form of equipment, or more generally, affordances. Importantly, such coping proceeds without a need for the agent to thematise the affordances that it is using, let alone the material things that anchor these affordances.

Our agency reading of Heidegger's phenomenological analysis of *Dasein* confronts that analysis with two questions that are not addressed in *Being and Time*, but which must be answered if the analysis is to be applicable to agency more broadly, and especially to artificial agency. A first issue is how to interpret the sensible agential possibilities that sustain an agent's coping with the environment: how are the existential possibilities for an agent anchored in the world as described by science? Secondly, it is striking that Heidegger's analysis does not address the becoming of the being that is *Dasein*—development and learning are mostly absent from his account. In view of our claim that agency is fundamentally characterised via learning, it is important to close that lacuna as well. We claim that both issues can indeed be successfully addressed in terms of our general account of agency (Chap. 2) and of the agency model of Projective Simulation (Chap. 5).

We bring up a phenomenological approach to agency at the end of our book not just for its own sake, but for three specific reasons that pertain to our overall project of finding a place for agency in our natural world. First, we started by pointing to agency as familiar from our individual experience as human agents; so it is fitting

to close with a view of agency that was developed with a specific view to *Dasein*, the human agent. Second, Dreyfus's critique of AI is a challenge for our claim, defended throughout the book, that artificial agency, which is arguably the most promising route towards Artificial General Intelligence, is a conceptual possibility, and that artificial agency can be modelled. Can the fundamental structures of agency coming from phenomenological analysis be recovered through science-inspired modelling? This question leads to the third reason for discussing phenomenology: With a view to the specific agency model that we discuss in detail in Chap. 5, Projective Simulation, we can ask whether *its* basic structures—its specific memory organisation, the decision finding process, and the learning rule—allow for a reading in phenomenological terms. As we will show, that is indeed the case. In this way, our discussion of the phenomenological analysis of agency is fully integrated with our explication of agency in Chap. 2.

References

- Bacon, F. (1620). *Novum organum*. John Bill. English translation as part of *The Oxford Francis Bacon*, Volume XII: *The instauratio magna: Part II. Novum organum*, edited by G. Rees and M. Wakely. Oxford University Press 2007.
- DiNardo, J. (2016). Natural experiments and quasi-natural experiments. In M. Vernengo, E.P. Caldentey & B.J.J. Rosser (Eds.), *The new Palgrave dictionary of economics* (pp. 1–12). Palgrave Macmillan.
- Dretske, F. (1994). If you can't make one, you don't know how it works. *Midwest Studies in Philosophy*, 19, 468–482.
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. The MIT Press.
- Hacking, I. (1983). *Representing and intervening*. Cambridge University Press.
- Heidegger, M. (1927). *Sein und Zeit*. Max Niemeyer. Quotations are from the English translation by John Macquarrie and Edward Robinson, *Being and Time*, Blackwell, Oxford 1962. Page references are to this translation, followed by the page number from the 19th German edition (Tübingen, Max Niemeyer Verlag, 2006) in brackets.
- Lear, J. (1984). The disappearing “we”. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 58, 219–258.
- Mulder, J. M. (2021). Animals run about the world in all sorts of paths: Varieties of indeterminism. *Synthese*, 199, 11945–11961.
- Nagel, T. (1986). *The view from nowhere*. Oxford University Press.
- Pearl, J. (2002). Reasoning with cause and effect. *AI Magazine*, 23(1), 95–111.
- Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.
- Pesic, P. (1999). Wrestling with Proteus: Francis Bacon and the “torture” of nature. *Isis*, 90(1), 84–94.
- Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
- Velleman, J. D. (1992). What happens when someone acts? *Mind*, 101(403), 461–481.
- Wheeler, M. (2005). *Reconstructing the cognitive world*. MIT Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 2

Agency



Agency is utterly familiar, yet deeply enigmatic. We and many other animals can do things, we have an influence on what happens that is due to us—and yet, we are part of nature just like anything else, material beings constituted by the same kind of matter and subject to the same natural laws. How do we fit in? The problem of agency, which we sketched in the introduction, is at root a conceptual one. It is not about the specifics of our own human agency, or about other biological forms of agency in the world, but about the notion of agency itself in its relation to and in contrast with the concept of natural happenings. The exact form of the agency problem depends on the specifics of how agency is spelled out. It is important, therefore, to clarify the notion of agency right at the start of this book. Our task is to capture a notion that is rich enough to sustain the special roles that agency and action play in contrast to related lower-level concepts such as events and causation, yet not so demanding that its proper application to a broad enough range of cases is called into question. One worry here is that our own, human case provides the most familiar example of agency, and so we may start modelling agency on our example—but we are not just agents, but also moral agents, and that higher-level concept must be kept distinct if we want to do justice to the fact that many animals act even though they are not moral agents, and that there could also be artificial agents that are not moral agents either. Morality does not form part of the notion of agency that we are after. A second, even more serious worry is that the demands on a notion of agency might turn out to be so high that the notion implies supernatural powers or even becomes incoherent. And a supernatural or incoherent notion could, of course, not be instantiated in our natural world, on conceptual grounds. This means, in particular, that the notion of agency has to fit into the natural order of things, despite the special autonomy vis-à-vis natural happenings that characterises agents. So, third, agency must not dissolve into mere natural happenings either. The notion has to avoid both extremes: a supernatural conception of agency, and the disappearance of the agent behind what happens anyway. On both these extremes, the specific phenomenon of agency could not be instantiated in our world, and we

would exclude the proper attribution of agency to machines, to animals, and even to ourselves.

Such an exclusion will not do, of course, because agency is a real phenomenon. Basic facts of everyday life like locomotion, ingestion and communication prove the existence of agency; there *really are* agents. A fortiori, agency must be possible: it must be possible that there is a special class of entities, agents, that are really different from other material beings. The intriguing issue left to discuss would seem to be *how* agency is possible, how it is that we (and other animals) are agents in this world. Answering this question through modelling might perhaps even pave the way for the implementation of agency in artificial systems. This will indeed be a main line of our investigation, but we have to address a number of philosophical worries about agency first.

2.1 Explicating Agency

No matter the philosophical position one takes up in the end, we have to start from a basic fact: We *have* the notion of agency, and we are highly skilled at employing it. We routinely distinguish agents from other objects, and actions from other happenings. This makes good evolutionary sense, given “[t]he high cost of failing to detect agents and the low cost of wrongly detecting them”: agents can pose dangers that natural circumstances don’t, and it is better to be safe than sorry. In fact, according to some theories, “people possess a Hyperactive Agent Detection Device, a cognitive module that readily ascribes events in the environment to the behavior of agents” (Gray & Wegner 2010).¹ We depend on agency detection not just for our safety—we rely on a robust notion of agency also for a host of our social practices as social animals. An action is something the agent is, in some sense, responsible for; a mere happening is not. When building trust in others or deciding how to reciprocate to what they do, the assessment of who did what is crucial. Thus, in everyday contexts we have no trouble deciding whether it was an action or, as we will say, an accident (a mere happening) that somebody stepped on our foot in a crowded bus, and to react accordingly. For philosophical discussions and for our purposes in this book, however, our intuitive everyday grasp of the notion of agency is not enough. Especially with a view to discussing models for agency and the possible agency of technological artefacts, it is crucial to work from a stable and clear base. So we need to clarify a notion of agency that is useful also in such contexts, confronting it with various philosophical challenges that we will unfold in what follows.

The philosophical method we will employ in our quest for clarification is that of explication, in which, broadly speaking, a concept is precisified starting from our

¹ This suggests, plausibly, that we employ the notion of agency also in cases in which (as we find out later) no agents were involved. We will discuss agency attribution in Sect. 2.3.2, and we specifically target cases in which it goes wrong in Sect. 2.4.

actual practices in which it plays a role, which includes an overview of intuitively clear cases of examples and counterexamples. The resulting iterative process of successive refinement² will lead to a concept that is better grounded and that provides more useful guidance in borderline cases than the intuitively handled concept that we started from. This may end up even changing the verdict on some of the initial clear cases. Such a shake-up happened, for example, with the notion of fish, which used to include whales.³

2.1.1 A Preliminary Account of Agency

We just said that we *have* the notion of agency, and that is true. But there is a lot of disagreement about that notion, about what exactly it is that we have. Or perhaps we should say that there are different notions that can all lay some claim to being “the” notion of agency. These span a wide range—from really minimal accounts that include software agents and according to which basically anything that effects a change (including a thermostat) is an agent, to highly demanding accounts, according to which only beings with a language and with morality—thus, we human beings—can be agents.

In a sense, there is nothing wrong with that diversity of course. “Agent”, unlike “Aronszajn tree”, is not a technical term with a precise definition, but a common word that has also been adapted to a number of more technical uses. It is just important to keep the various uses distinct. In this book, we target a notion of agency that is not so demanding as to exclude all non-human animals, and yet not so broad that it includes simple mechanisms. When we use the term “agent” or “agency”, we mean *that* notion. This does not mean that other uses of the term are improper or bad; it is just that in this book, we label them explicitly to mark that distinction, so that we avoid ambiguity. Thus, we will speak of “moral agents” or “linguistic agents” when we consider a notion that is so narrow that it applies only to us humans, and we will speak of “percept-action agents”, or “reflex agents”, or “software agents” when we consider one of the much wider notions employed, e.g., in computer science or in robotics (Russell & Norvig 2022). Our target notion of agency is, as we will argue, ontologically relevant in that it marks a certain level of organisation that is really and distinctly exemplified in our world. That notion is applicable to many, but not to all animals, marking a relevant distinction within the animal kingdom. As a first approximation, bacteria and other simple organisms are excluded, while vertebrates and some invertebrates like the octopus are included.

² Another term in use for this process is “conceptual modelling” (Löwe & Müller 2011). Compare also the “connective model” of analysis described by Strawson (1992, 21).

³ In Melville’s *Moby-Dick* (§32), Ishmael classifies fish by the orientation of their tail fin: that of the whale extends horizontally. But we all know, of course, that whales are not fish, but mammals. Cladistically (considering evolutionary ancestry), however, it turns out that whales and all other mammals, indeed all tetrapods *are* fish. This mess is one of the reasons why the concept *fish* is no longer considered useful for biological taxonomy nowadays.

There is not a lot of philosophical literature that helps with the delineation of agency in our sense; much of the philosophical discussion revolves around the human case.⁴ Perhaps the closest ally, and certainly a philosopher from whose work we have profited a lot, is Helen Steward, who discusses agency, explicitly including animal agency, in her book *A Metaphysics for Freedom* (Steward 2012). She provides a list of criteria that we find highly useful. Our own approach will be different from hers, as we will discuss below: we place emphasis on learning and memory, which is not included in Steward's list, and we strive to stay away from mentalistic vocabulary even more than she does. These changes of focus are due, at least in part, to our interests in the diachronic aspects of agency and in agency modelling. We believe that much about agency (and perhaps also about the possibility of artificial agents) can be learned when one takes the historicity of agents seriously. As we will argue, it is significant that all agents that we know develop over time, building up a memory, adapting to their environment, and thereby acquiring individual traits. But before we get into all these details, it will be useful to have a look at Steward's list of criteria of agency.

According to her definition, agency is characterised by the following four aspects, which are all necessary, but not strictly independent, and which should therefore be viewed as a package (Steward 2012, 71f., italics in original):

- i. an agent can move the whole, or at least some parts, of something we are inclined to think of as *its* body;
- ii. an agent is a centre of some form of subjectivity;
- iii. an agent is something to which at least some rudimentary types of intentional state (e.g. trying, wanting, perceiving) may be properly attributed;
- iv. an agent is a settler of matters concerning certain of the movements of its own body [...], i.e. the actions by means of which those movements are effected cannot be regarded merely as the inevitable consequences of what has gone before.

We comment on this list briefly, so as to explain some of its features, and also with a view to a later comparison with our own list, which we present in Sect. 2.6.

i. A basic truth about agency is that agents make a causal difference to what happens in the world. This means that they need to be embodied: agents are concrete

⁴ Agency is currently becoming more visible as a philosophical and as a scientific topic—witness, for example, Tomasello (2022) and Mitchell (2023), and see Ferrero (2022) for a broad philosophical overview. Ferrero proposes to shelve the old labels of “action theory” and “philosophy of action” as too narrow (p. 15f.), and we recommend both his introduction to the volume as well as its individual articles. The volume does, however, not provide a definition of the notion of agency. Ferrero gives good reasons for this abstention and instead charts four different *pictures* of agency that inform various parts of the philosophical discussion (p. 8ff.). From our point of view, what stands in the way of a more focused definition is the—fully justified—observation that philosophers discussing agency often have a focus on, or at least a special interest in, what Ferrero calls “full-blooded agency” (p. 7f.). This notion is human-specific and more or less coincides with what we call moral agency. For our purposes, which include providing a target for agency modelling, we develop a notion of agency specifically *below* that level.

material beings.⁵ Steward's item (i) is usefully more specific in that it makes more precise the kind of embodiment in question. Agents *have* a body: they are not simply identical to their bodies, but have some control over them, in the sense of a capability for self-movement. Steward is careful and allows for the possibility of limited self-movement (as in sessile barnacles, who in adult life can only move their limbs), but she also makes it clear that normally, agential self-movement implies locomotion, i.e., the ability of the agent to move its whole body through space.⁶ Steward hedges her item (i) with an epistemic clause that, given our focus on the ontology of agency, we find superfluous. Whether "we are inclined to think" this or that should not play a role when it comes to the question of whether some entity is an agent or not. But she is right that we *do* perceive self-movement, and we do make distinctions between entities who are identical with their bodies and entities (agents) who *have* a body. Furthermore, it is certainly no coincidence that when facing an agent, we are inclined to think that we *are* facing an agent.

ii. Steward's second item connects with what we find the most difficult aspect of agency to make precise: an agent is or has, in some sense, "a self". What exactly this means is not easy to spell out. Steward briefly discusses the question of animal consciousness (Steward 2012, 72), and without committing to philosophically controversial claims, points out that agency comes with at least some notion of a *perspective* on the world; a view from somewhere, as it were. We basically agree; we just try to limit the mentalistic aspects of our criterion even more than she does. As we will lay out, the notion of an agent's learning history will be helpful in that respect, as it allows us to anchor an agent's individuality, and thus its individual self, in its history of interactions in dealing with the environment, which is embodied in the agent's memory. We will, however, also need to take recourse to the idea that there has got to be something "in it" for the agent with respect to what it does, and this is quite close to ascribing some form of subjectivity (see Sect. 2.6 for discussion).

iii. With her third item on the list, Steward points to an aspect of agency that may be closely connected to item (ii), namely, that agents want certain things (but not others), that they try things, and that they perceive their environment in some way. She ties these aspects to the notion of intentionality, which at root means to be mentally directed towards something.⁷ She is, however, careful to avoid

⁵ "Concrete material" here contrasts with "abstract" and with "immaterial". Agents are not abstract entities like numbers, they are concrete individuals. Furthermore, while we acknowledge that in philosophy of religion, there are discussions of immaterial (e.g., divine) agency, we leave that debate to the respective experts and consider only what connects well with our understanding of physics. "Material" in this sense is to be taken broadly: for example, photons are material in our sense.

⁶ For example, in barnacles, the larvae are able to self-move in this way. None of this is meant to imply that the barnacle is an agent; we can leave that issue open for further scientific research.

⁷ Brentano famously took intentionality, understood as mental directedness, to be the defining feature of mental phenomena (see Brentano 1874, 92f.). As we will point out in Sect. 2.6.1.iii,

commitment to typical explications of intentionality in terms of beliefs and desires in which the latter notions are taken to be propositional attitudes: that approach threatens to limit intentions to beings with a language, i.e., to us.⁸ We are firmly in line with Steward's cautionary approach here. It is clear that agents must be able to make distinctions in their interactions with their environment, which must be based both on some access to the environment (perception) and some form of wanting something for their own sake. But these dynamics need not be spelled out in the form of logical inferences of the type that are sometimes taken as definitive of intentionality (witness the influential "belief-desire-intention" model of agency, inspired by Bratman 1987): "I want it to be the case that *X*, I believe that doing *Y* is the best means for making it the case that *X*, so I do *Y*". Such inferences can be useful, and we sometimes have recourse to them in explicit planning or in specific decision situations, but most of the time we act without them, and beings who cannot draw such inferences may still be agents in the literal sense of the term (see our discussion of circumspect coping in Sect. 6.5.3). We find it useful to speak of "wants" rather than "intentions", but only to further stress that, like Steward, we associate agency with intentional states of only a rudimentary form. Neither having a language nor having the capability of drawing logical inferences is presupposed.

iv. Steward's item (iv) expresses her conviction, which we share, that agency needs indeterminism and is therefore incompatible with determinism. (This is the reason why she calls her position "agency incompatibilism"; see also Steward 2020, 2022.) An agent is, in her words (echoing Belnap et al. 2001), a *settler* of matters concerning its self-movement. This means that from an available range of really possible courses of events, an agent has the capability of making a difference as to which of these possibilities is realised. Settling in the technical sense implies that before the action, a range of options was present, while after the action, one of these options has been realised, and that realisation was *due to the agent*. Thus, metaphysically, the notion of an open future is fundamental for agency, and agency presupposes freedom to influence the future course of events: as Steward says, an action "cannot be regarded merely as the inevitable consequences of what has gone before" (Steward 2012, 72). Again, we believe that the epistemic hedging ("cannot be regarded") is superfluous; what is important is that the action is not

we try to stick to the terminology of *wanting* so as to avoid the strongly mentalistic connotations of intentionality. See also Sect. 6.2 for further discussion.

⁸ A proposition is the content of a thought or a judgement, like "The window is closed", which can also be the content of a question ("Is the window closed?") or of a wish or a desire ("I want it to be the case that the window is closed"). The wish has the opposite "direction of fit" (see Humberstone 1992) from the judgement: a judgement is true if its content fits with what the world is like; a wish is fulfilled if the world fits its content. These various ways of making use of a proposition are called *propositional attitudes*. Propositions are compositionally structured entities, so that having propositional attitudes basically requires one to have linguistic capabilities.—In line with Steward's approach, we focus on a less demanding notion of wanting, according to which wants do not need to be propositionally structured. We are also open towards a correspondingly less demanding understanding of beliefs.

pre-determined, and that agency is incompatible with determinism. Steward agrees with this, and she provides a detailed argument for her agency-incompatibilism in the course of her book (Steward 2012, Ch. 5–8). Her discussion is closely connected with arguments given in the free will debate, which she (in our view, absolutely correctly) identifies as being to a large extent not about free will in the morally loaded sense, but about the more basic notion of agency. For the purposes of this book, we stay even further away from the free will debate with its focus on moral responsibility. By focusing on the ontological question of what an agent is, we hope to be able to provide a complementary argument for agency incompatibilism that strengthens the points that Steward makes: under determinism, there could be no actions because *there could not even be agents*.

2.1.2 Topics to Be Addressed in This Chapter

The bulk of this chapter will be devoted to working towards our refined explication of agency, which we will provide in Sect. 2.6. As we just mentioned, Steward’s list provides some helpful guidelines for us, but our focus is also quite different. One of our aims is to stress the ontological reality of agency vis-à-vis non-agentive nature, and for that we need to prepare the ground. Thus, in Sect. 2.2, we discuss basic notions of ontology, of what there is. We point out the dialectics of reductionism: once we see a larger entity (such as an agent) as composed of smaller bits, we may start to wonder whether we really need the higher-level category. Doesn’t the microphysics say it all? After our discussion of forms and strategies of reductionism in Sect. 2.2.2, we provide an argument for the reality of ontological levels in Sect. 2.2.3. Our claim is that there can be ontologically emergent individuals, which are *not* mere aggregates of their constituent matter, but whose organisation uses that matter as material, e.g., to maintain a dynamic equilibrium. We illustrate this idea with the example of the sun, thus defending the claim that the sun is really real. It is crucial for our argument that the lower-level matter afford enough *open possibilities* such that a higher-level entity can make a difference through organising that matter as its material. If lower-level constituents were subject to deterministic dynamics, no higher-level ontology would be possible. This is, briefly, how we complement Steward’s argument for agency incompatibilism through our focus on ontologically emergent individuals.

Having laid the groundwork, we go on, in Sect. 2.3, to discuss the agency question as a question of ontology. We discuss the role that agency plays in some of our basic practices, namely in language (Sect. 2.3.1) and in agency attribution (Sect. 2.3.2). In the latter discussion we propose a “ladder of attribution” on which agency attribution is sandwiched between a lower rung of mere causal attribution and a higher rung of moral attribution. In further examples, we bring to the fore the important role that diachronic considerations of an agent’s history, memory, and training play in these attributions (Sect. 2.3.3). We complement these discussions with an in-depth treatment of defeaters of agency attribution, looking at cases in

which we retract a previous attribution of agency due to new information, e.g., about the foreign control of a would-be agent (Sect. 2.4). Before the background of this overview, we then consider the possibility of artificial agents, and thus, of agency in non-biological systems. We affirm this possibility in Sect. 2.5. These discussions, on the one hand, affirm our realistic attitude towards agency: agency is real, and attributing agency is something one can get right or wrong; it makes a real difference. On the other hand, the discussions provide material for our novel explication of the notion of agency, which we spell out in Sect. 2.6. This provides a target notion for agency modelling as projective simulation, as we argue in Sect. 2.7.1. The chapter closes with a brief discussion of the role of agency in science (Sect. 2.7), which already prepares for some of the discussion to follow, starting with quantum mechanics in Chap. 3.

2.2 Ontological Levels of Being

The problem of agency that we introduced with Nagel's worry was this: agency has to fit into the order of nature—but then it will be of a piece with the rest of nature, and thus cannot be so special after all. Never mind that we find good uses for a notion of agency in the interest of our survival and of our social arrangements; when we consider the matter scientifically, agency dissolves into patterns of natural happenings. Does this mean that, *really*, there are no agents? This worry concerns a question about ontology, about what there is. Are there agents? Well, we are agents, so it follows that there are agents. That is a valid inference. But does it inform us about ontology, about what there *really* is? Questions about what there is are curiously tricky. As Quine famously said, the answer to the question “What is there?” is simply “Everything”.⁹ But that answer by itself is, of course, not helpful.

2.2.1 *Ontology and What There Is*

Questions of ontology need to be separated from mere questions of existence. Not all questions about what there is are questions that pertain to the philosophical discipline of ontology. In fact, most are not. Questions such as “Is there a prime number between 1000 and 1010”, “Is there life on Mars”, or “Is there a fair way to elect a government” are normal scientific questions of mathematics, biology, and political science. Ontological questions operate at a higher, more abstract level. They pertain to categories of being rather than to individual instances of these

⁹ See the classic opening of his essay “On what there is” (Quine 1948): “A curious thing about the ontological problem is its simplicity. It can be put in three Anglo-Saxon monosyllables: “What is there?” It can be answered, moreover, in a word—“Everything”—and everyone will accept this answer as true.”

categories: Are there numbers (in addition to our practices of talking about them)? Is there life (in addition to the physical and chemical processes that go on in living beings)? Are there states and governments (in addition to the people that make them up)? Ground-level scientific claims normally do not have an impact on these higher-level issues, which are more closely connected to questions of the form “What is an *X*?” than to “Is *a* an *X*?”. The fact that 7 is a prime number is, first of all, a fact that is grounded in our practice of dealing with numbers; it does not show that there are numbers *in addition to our practices of talking about them*. More work (e.g., on what numbers *are*) seems needed in order to pass a verdict on that. This means that the task of establishing an ontological category is a specifically philosophical one.

Why not simply accept that there *really are* agents, as our practices of identifying them seem to imply—why worry about ontological categories? The slogan of Occam’s razor identifies a valid concern: Entities should not be multiplied beyond necessity (or: a plurality should not be posited without necessity).¹⁰

There is certainly a sensible presumption against multiplying categories of being. Our account of what there is should not be needlessly complicated. The last two words of the slogan are, however, important: it is not at all forbidden to have more than one ontological category. But each introduction of a separate category has to be necessary. That, of course, leads to the next hairy question: what is necessary?

Taken lightly, one might say that is a matter of taste. Quine admitted of a liking for “desert landscapes”, for a minimalist ontology that only admits what is absolutely necessary in the interest of natural science (cf. Quine 1948, 23). And certainly there are electrons, protons, neutrons, and such things; do we really need to assume that there is more, isn’t everything else just an aggregate of those basic entities? Many feel the pull of that question and are inclined to favour ontological minimalism. But while perhaps a majority of philosophers shares some version of Quine’s preferences, others—again in the interest of science—favour a more colourful and disorderly view of nature that admits a diverse set of ontological categories (e.g., Dupré 1995; Cartwright 1999).

Do we have to leave the matter at the level of taste? The fact that the mentioned two opposing camps both refer to science raises hopes that arguments can be given, not just expressions of personal predilections. But as so often in philosophy, the arguments are complex. In this case, they are also entangled with the respective views of what science is and how the various sciences that we have (in the broad sense of *Wissenschaften*, which includes formal and social sciences and the humanities) are related. Here, the minimalist camp favours the view that the division of the sciences into many different ones is only a temporary glitch, to be amended by a future unified science built on solid foundations.¹¹ This was in fact one of

¹⁰ According to Cloeren (1984), the most common formulation in Occam is “Pluralitas non est ponenda sine necessitate”, while the commonly attributed phrase “Entia non sunt multiplicanda praeter necessitatem” cannot be found in Occam’s own writings.

¹¹ Historically, it has not at all been clear what these foundations should be. The hope for indubitable empiricist foundations has pushed some towards phenomenological entities: the basic elements on which all of science should be erected would thus be individual sense impressions.

the key projects of logical empiricism, the main historical precursor of present-day philosophy of science: their programmatic book series that was launched in 1938 was called the *International Encyclopedia of Unified Science*.¹² The other camp, by way of contrast, affirms the separate methodologies and subject matters of the different sciences and argues that while the divisions are historically contingent and fluid, and interconnections are common and important, a division of the sciences can be based on real distinctions in nature.

Those who favour a unified science (and, by way of Occam's razor, a minimalist ontology) may argue something like this. Never mind the formal sciences of logic and mathematics. And let us not worry about the humanities, history and all that—in some sense, these are social sciences. Looking at the natural and the social sciences, we can acknowledge that these study the world basically at different length scales or, which more or less amounts to the same, at different levels of organisational complexity. After all, larger things consist of smaller things. And the smallest things we know of are elementary particles, which are studied by physics. Anything at all—a rock, a flower, an octopus, a city, a planet—ultimately consists of these smallest entities. So if we understand the properties of these entities and the ways in which they combine, we have understood all there is to be understood about the world. We may need to take some shortcuts in order to work out specifics, because of our perceptual and cognitive limitations: we cannot work out the interaction of a group of people in a bar at the level of elementary particles, because the computations are too hard for us. But surely anything that happens in the bar comes down, ultimately, to the interaction of these elementary particles.

This picture has proven to be seductive; in fact, it appears to inform a large part of the current perception of science in the public. Chemistry is just applied physics; biology is just applied chemistry, and so on. Divide up a flower into smaller parts, you find its organs, made up of cells, made up of molecules, made up of atoms, made up of elementary particles. Physics wins out; philosophically, we are driven towards physicalism, the ontological assumption that all there *really* is, is the fundamental entities investigated by physics. This picture is reinforced by a reflection on the historical contingencies of the division of the sciences: many divisions have arisen only relatively recently, institutional pressures have played a role, and there are many in-between fields, from physical chemistry to social psychology. From this

That project goes in the direction of idealism; it has never been worked out in enough detail to be confronted with scientific practice. Other foundationalist programs build upon ordinary objects (Austin's (1962, 8) "moderate-sized specimens of dry goods") or some set of fundamental entities of natural science, e.g., elementary particles and/or some fundamental physical fields. In this case, it has to be acknowledged that the foundations themselves may be open to revision, just like the rest of science. See Uebel (2007) for an overview of the 1930s exchanges on protocol sentences in the Vienna Circle, where this issue was hotly debated.

¹² See Pombo (2011) for an overview. The project was never completed. Perhaps ironically, Kuhn's *Structure of Scientific Revolutions* (1962), a main force behind philosophy of science's turn to history and on to the more detailed study of actual scientific practice, was one of the last volumes in the series.

point of view, it looks more like a continuum divided up along contingent boundaries than a proper partitioning of nature “at the joints”.¹³ And the basis on which it all rests is the study of the smallest things we find in nature, elementary particles or whatever there is at the most fundamental level. Anyway, the smaller, the more fundamental.

Many scientists would probably agree to some version of the above claim. Metaphysicians are more divided—it is even philosophically controversial whether there really is a fundamental level of reality.¹⁴ But the more important worry is that the unificationist project has consistently failed to deliver. Worse still, as we will argue in Sect. 2.2.3 below, it appears to rest on a dubious assumption of determinism, which is rejected by our current candidate for the lowest-level physics, quantum mechanics. With respect to scientific arguments for or against a minimalist ontology, we thus have an opposition in which the one side can refer to an established and, for all its imperfections, functioning pluralism, while the other side envisages a grand reform in the interest of some philosophical ideal. It appears to us that the burden of proof lies with those who affirm the unity of science and argue for a minimalist ontology on that basis. But, again as so often in philosophy, the matter is controversial. And even if some form of pluralism wins the day, this does not settle the matter of exactly *which* ontological categories make up the plurality of what there really is. Case in point: Pluralism alone is not sufficient to prove the reality of agents as a distinct ontological category.

2.2.2 *Reductionism: To Eliminate or not to Eliminate?*

The drive to accept as few ontological categories as possible may not lead to just a single category, but it will probably lead to the exclusion of some would-be categories. And whenever an aspiring category is excluded, that necessitates a reductive treatment of what appears to rely on it. Even if we exclude a category from the list of what there *really* is, the fact remains that we will have practices that appear to rely on that category, and these practices need to be dealt with as well. Even if there *really* are no numbers, you still want to get the correct change at the supermarket till, and even if there *really* is no separate ontological category of living beings, you wouldn’t want to buy the dead parrot in Monty Python’s pet shop.

A reductive account thus has to deliver two things, which often come packaged up together. First, there has to be some account of our talk in terms of the to-be-reduced category, or more generally, of our practices in which that category is involved. Second, there has to be an account of what is left of the entities that have been reduced away, ontologically speaking.

¹³ The image is that when studying nature, one should proceed like a good butcher, who carves at the joints, where the natural divisions are. See Plato’s *Phaedrus* 265e.

¹⁴ See, e.g., Ladyman and Ross (2007) for a view that is generally critical of the idea of low-level entities, championing ontic structural realism instead.

For our purposes, it will not be necessary to delve into the many fine distinctions that can be made at this point.¹⁵ There is one distinction, however, that is crucially important both in general and with a view to our discussion of agency in particular: the one between eliminative and non-eliminative reductions.

Eliminative Reduction In philosophy, eliminative reductive strategies have been proposed, e.g., for abstract entities such as properties and numbers, for mental categories, or for moral values. When one eliminates, one says that there is *nothing*, really, that corresponds to the would-be category in question. Now eliminativists about numbers may prune away all abstract entities from their list of ontological categories, but they still have to come up with some story for why we hold “ $1 + 1 = 2$ ” to be true and “ $1 + 1 = 3$ ” to be false—even though, strictly speaking, these statements cannot be read as assertions of identity of numbers, as there are no such things. Various strategies have been devised in this specific case, for example, (neo-)logicism, which accounts for these and other statements in terms of logical formulae.¹⁶ Other strategies have to be used for other cases of elimination. For example, one can hold an error theory that claims that the statements in question are all really false or even meaningless, but have some pragmatic justification, e.g., as expressing certain emotions.¹⁷

Non-eliminative Reduction Eliminative reduction is a radical claim. Reductive accounts in the realm of the natural sciences, by way of contrast, are normally non-eliminativist. At least we do not know of anybody who has seriously proposed, e.g., eliminativism about chemical categories. Still, reductionism about chemistry, biology and so forth is common, as we just pointed out. A non-eliminativist reductionist will not say that terms such as “hydrochloric acid” do not really refer to anything real, but they will suggest that there is some way of doing away with these terms nonetheless. Thus, a statement such as “If you combine hydrochloric acid and sodium hydroxide, you get table salt and water” is still taken to be literally true, but it is claimed that everything that is said in this statement can equally well be expressed via a statement in purely physical terms, which shows that the statement rests on nothing but physical facts. (Never mind that no explicit full reformulations are ever written down.) For an intra-physical case, consider the assertion that “Water is a fluid at room temperature”. That is of course so; and it is surprising because

¹⁵ Most importantly, we will not be concerned with reduction as a relation between scientific theories, which is much discussed in philosophy of science. See Van Riel and Van Gulick (2025) for a comprehensive overview.

¹⁶ Frege defended logicist foundations for mathematics in his *Grundgesetze der Arithmetik* (Frege 1893). Frege’s own approach was found to be inconsistent: Russell formulated what is now known as the Russell paradox (Is the class that has as members exactly those classes that are not members of themselves, a member of itself or not?) to point out a problem with Frege’s Basic Law V that could not be repaired. New and improved versions of the original idea of logicism have, however, since been worked out. See, e.g., Tennant (2023) for an overview of neo-logicism, which also includes a discussion of the Russell paradox (§1.2.5).

¹⁷ See Mackie (1977) for the *locus classicus* of an error theory about ethics.

water molecules are so light that one would expect that, like carbon dioxide for example, water should rather be a gas. Here it is perhaps plausible to say that the fluidity of water at room temperature is *nothing but* the effect of hydrogen bonding between the water molecules, which is made possible by their bent geometry. In making that claim, the non-eliminativist reductionist does not deny that water is a fluid, but claims to account for that in terms that do away with fluids as separate entities. Thus, to the question of what *really* is behind a would-be category, the non-eliminativist answers not “nothing”, which is the answer given by the eliminativist reductionist, but “nothing but . . .”. Chemical substances are nothing but aggregates of physical entities; living beings are nothing but chemical aggregates; and so on.

Reductionism about life thus also normally takes the form of non-eliminativism: The attempt is to account for the vital categories of organs, development, and life processes generally in terms of nothing but biochemistry, which in turn is then additionally often taken to come down to nothing but physics. The attempt is not without its serious problems,¹⁸ but it appears to be almost a part of being scientifically minded these days. On the other hand, eliminative reductionism about life, which would claim that there is, really, no life on earth, seems wholly unwarranted.

Reductionism About Agency? The picture is much less clear when it comes to agency. On the one hand, this has to do with the fact that agency as a separate ontological category is little discussed, and that many discussions of agency are phrased in terms of the more complex notion of (human) *moral agency*, for which the contested ontological status of moral values and concerns about the notion of free will create additional difficulties. But there is also another issue: Some discussions of agency take the notion to be ultimately contradictory or incoherent, perhaps a remnant of some form of pre-modern superstition. That would of course mean that we have to be eliminativists about agency, since that which is contradictory or merely an element of superstition cannot be instantiated in our natural world.

Doing Away with Superstition Some eliminativist reductions are indeed part of the enlightenment and have stuck with us, let us hope, for good. There are, for example, no witches, and historical practices of witch hunting and so forth have been unmasked as morally repulsive ways of mistreating, torturing, and killing thousands of human beings (mostly women). With monarchies and other forms of hereditary superiority, the record is perhaps less firmly settled; Russell famously said that the British monarchy had been allowed to remain “only because it is erroneously supposed to do no harm”. The essay from which this quote is taken, “On the notion of cause” (Russell 1912), was his attempt at eliminating not the monarchy, but causation, which Russell also held to be “a relic of a bygone age” (Russell 1912, 1) that was scientifically fruitless. While Russell’s attempt at elimination looked to

¹⁸ See Mulder (2016) for a succinct statement of problems with the position. Mulder argues convincingly that these problems cannot be overcome.

be quite successful for a while,¹⁹ the pendulum appears to have swung back by now, and causal talk is taken to be philosophically and scientifically respectable in most quarters nowadays. This is certainly, at least in part, due to the importance of causal analysis and causal learning as a major field of data analysis, machine learning, and AI in general.²⁰

Russell's argument against causation was that the notion has no proper application in the advanced sciences, so that it was unnecessary and should not be allowed to linger around. For some other notions, the drive towards elimination has been motivated by the attempt to show not just their bad ancestry or their uselessness, but their incoherence. Thus, for example, Spinoza and Voltaire argued that the notion of a miracle was based on a conceptual contradiction, from which it would follow directly that there can be no miracles, and Hume famously argued that the notion of a miracle was such as to rule out that evidence of a miracle could ever be rationally believed.²¹ Whatever the merits of Hume's argument, some practices involving miracles are still with us.²²

What About Agency? Is agency a relic of a bygone age? At least there are some arguments to the effect that the notion relies on something akin to miracles or is incoherent and should therefore be abandoned—which would motivate eliminative reductionism regarding agency. Overtly, most of the arguments of that type are directed not at the notion of agency itself, but at the notion of free will (e.g., Pereboom 2001, 2014)—but by their structure, they often target agency, not the higher-level notion of *moral* agency that is traditionally connected with free will. On the view of agency targeted by some arguments, an exercise of free agency would have to contravene the laws of nature, and thus be very much like a miracle according to Voltaire, or at least would presuppose, echoing Hume, “coincidences too wild to be credible” (Pereboom 2014, 66). We will discuss the general form of such arguments under the heading of “agency scepticism” in Sect. 2.4 below. In our view, these arguments are not successful. It is true that agency presupposes a certain kind of autonomy with respect to natural happenings: Agents can make things happen that would not have happened if it were not for them. Rabbit holes, bird's nests, and cities do not come about by accident via non-agentive natural processes such as the motions of the tides, erosion, or sedimentation. And, closer to our main

¹⁹ For example, the logical empiricists of the 1930s and onward banned the notion of causation from serious talk as being metaphysical, and would only acknowledge practices of scientific explanation.

²⁰ Witness, e.g., Peters et al. (2017) or Pearl and Mackenzie (2018).

²¹ Hume's argument appears in Section 10 of his *Enquiry concerning Human Understanding* (Hume 1748). It has led to a large and controversial literature. See the overview by McGrew and Larmer (2024), which also contains other relevant references, e.g., to Spinoza's 1670 *Tractatus* and Voltaire's 1764 *Philosophical Dictionary*. See also Levine (2011) for an overview emphasising the point of view of philosophy of religion.

²² A miracle is needed, for example, in the official rules for canonization of the Catholic church, unless the person in question is a martyr. See Del Colle (2011) for a historical overview.

interests in this book, scientific experiments do not come about by chance either, but take a lot of agentive intervention, careful planning, and active shielding against outside influences. But this independence of agents vis-à-vis natural happenings should not be understood as contravening the laws of nature. Rather, agents can exploit the wiggle room left open by these laws to further their aims. As we will discuss, this would indeed not be possible if the laws of nature were deterministic. But they aren't—this is one of the important lessons of quantum physics. On the general picture of ontological categories as building up through “matter used as material” that we will discuss in the following Sect. 2.2.3, agency is just one specific example of how lower-level possibilities can be organised in the interest of higher-level structures.

We hold, therefore, that attempts at an eliminative reduction of the notion of agency are unsuccessful. But this still leaves open the second, non-eliminative version of reductionism. Agents clearly are material beings, ultimately constituted by physical particles—isn't it, then, that agency is nothing but some fairly complex physics, so that there is no need for acknowledging a separate ontological category of agents after all? One pertinent answer to this challenge is that we have not seen any plausible development of a reductive account of agency in physical terms, so we have good reasons to be sceptical. Of course, when one investigates an agent by physical or chemical means, one will find that it has normal physical and chemical properties—a mass, a volume, a chemical composition, and so on. But agents have other properties as well, and no plausible way of unpacking these into physical or chemical terms is anywhere on the horizon. To say that such unpacking has to be possible “in principle” is just to reiterate the reductionist sentiment, it does not amount to an argument. The burden of proof lies with those who would want to do away with the ontological category of agency that we have.

2.2.3 *A Positive Account of Separate Ontological Categories: Ontologically Emergent Individuals*

So far, our argument for holding on to agency as a distinct ontological category has been mainly negative. Quite generally, the dialectical situation that we have described seems to be this. If there are valid reasons showing that a would-be ontological category rests on inconsistent assumptions, or on mere superstition, then that level has to be reduced away (eliminated)—even if there is no alternative yet for dealing with all the practices in which that level is implicated. (The reduction can in fact usher in a change in our practices, some of which may consequently be abandoned. Such was the case, let us hope for good, with witch-hunting.) Otherwise, a challenge to an ontological level that is ingrained in our practices has to be supported by a *detailed* (not merely programmatic) account that shows how the notions connected to that level can be avoided or replaced. (For example, neo-logicism about mathematical objects is a serious attempt at providing such a detailed account; see note 16.) If a critique of an ontological level is not supported in one

of these ways, the burden of proof remains with those who mount the reductive challenge. By this count, then, the ontological levels of life and of agency certainly stand firm.

But a merely negative defence may still feel unsatisfactory, and there is a lingering general worry that threatens to tilt the dialectical stand-off in favour of the reductionist. Assume that the microphysical level is real—and it certainly is.²³ Now, how could other layers of reality not be *nothing but* their constituent atomic and sub-atomic parts? In which sense can there be more than what you find when you look as closely as you can? Isn't it simply like this: Take any material thing, divide it up, you find its chemical constituents, which in turn are constituted by the familiar elementary particles, interacting via the familiar fundamental physical forces. So it turns out that the thing *is made up out of* these constituents, connected by these forces. And doesn't this prove that this is all there is to the thing, and that—at least in principle—it must be possible to *construct* the thing from these constituents, by bringing them together and adjusting the forces between them in the right way?²⁴ Therefore, doesn't the proponent of a higher ontological level have to either assume non-physical constituents or forces, or give up? And the case for such forces or constituents, which perhaps was an open issue in early twentieth century discussions of biological principles like *entelechy*, is surely now closed. Doesn't this provide a *reductio* of the idea of higher ontological levels?

This worry does affect agency, life, and all other purported higher ontological levels of material things, at least in the sense that they are open to non-eliminative reduction, which threatens their autonomy. It would be beneficial if defenders of the

²³ There were justifiable doubts about the atomic nature of matter until the beginning of the twentieth century. But our ever increasing experimental capabilities of manipulating matter at the atomic scale show convincingly that *there are* atoms, protons, neutrons, electrons, and photons. In saying this we rely on Hacking's (1983, 23) "if you can spray them then they are real" argument for entity realism. Hacking was convinced of the existence of electrons by the experimentalists' capability of spraying single electrons onto oil drops in the Millikan experiment. Present powers of manipulation go much further. Witness, for example, recent experimental advances in the mechanical manipulation of single cells and biomolecules with the help of optical tweezers and atomic force microscopy (Neuman & Nagy 2008; Bustamante et al. 2021), or in the manipulation of both the external and internal degrees of freedom of individual atoms in large atomic arrays in quantum computing architectures (Bluvstein et al. 2024).

²⁴ Just think of how readily many of us take in the image of the Star Trek transporter, a device that disassembles and reassembles material objects, including living beings, to transport them over distances of several thousand kilometres. This idea lives off the aggregate picture of macroscopic entities—in fact, in the context of the imagined transporter technology, any material being, alive or not, is viewed as a three-dimensional bitmap that can be read out and, using the information read out, reassembled from its original or from other matter piece by piece. (See Sect. 3.4.2.2 for a more detailed description of the bitmap analogy.) Part of our work here is to show in which way this image is misleading.—We will discuss so-called quantum teleportation in Sect. 3.4.4.1 below. That protocol shows how quantum information, which cannot be copied, can still be transferred. Quantum teleportation can therefore not be understood in terms of read-out and reassembly. This supports our verdict against the bitmap image. In addition, note that the transfer of quantum information in quantum teleportation uses a previously established, shared material resource.

ontological reality of agency had more to offer than a reference to the dialectical situation concerning the burden of proof pointed out above. That is, they should ideally provide an explanation for *how* higher levels of material organisation can have ontological reality in addition to (or on top of) the microphysical.

This challenge has two aspects, one conceptual and one empirical, and these are inextricably intertwined. The conceptual issue is to provide a criterion of success: when will it have been shown that some phenomenon is a true witness of a higher ontological level? And the empirical issue, then, is to provide such a witness, or better, a class of such witnesses. As so often in philosophy, the trouble is that there are good arguments for different and conflicting criteria, and there may be the temptation to model one's criteria after an empirical witness that one already has in mind. So it is perhaps no big surprise that the whole area is highly controversial, and some of the exchanges even look as if all constructive communication between the conflicting parties has broken down.²⁵ Even the terminological framing of the conceptual issue offers room for disagreement.

2.2.3.1 Holism, Top-Down Causation, Emergence

The conceptual issue of providing criteria for the existence of a higher level is often discussed under the heading of one or more of the technical terms of *holism*, *top-down causation*, or *emergence*. Each of these notions comes with a certain image of what a higher level should be. Holism is connected to the ancient (Aristotelian) slogan that “the whole can be more than the sum of its parts”, but the notion has a vast array of different and partially conflicting uses, and so we will not focus on it here. The image of top-down causation is that an ontologically higher-level entity can causally influence its parts, which seems clear enough. And it seems perfectly obvious that there are cases of top-down causation in agency: for example, you can drink a glass of water or clip your fingernails, and thereby cause a change (via addition and subtraction, respectively) to your lower-level material constitution. But opponents of top-down causation will then argue that such homely examples beg the question as they presuppose our everyday (thus, unscientific) frame of mind, and they will also point out that top-down causation (including any instance of mental causation) must be rejected because it violates the alleged causal closure of the microphysical,²⁶ which proponents of top-down causation can then criticise as begging the question against them. Add to this the many controversial debates about the notion of causation itself,²⁷ especially on the contested interrelation of causation

²⁵ For some pertinent remarks along these lines, see, e.g., the 2014 introduction to the reprint of Anderson's classic paper “More is different” (Anderson 1972; Goldstein 2014).

²⁶ There is a large and controversial literature on the so-called causal exclusion argument, often with reference to Kim (1998). See Robb et al. (2023) for an overview. Like Kim, that overview focuses on mental causation, but the authors additionally provide a wealth of more general references (see esp. §6).

²⁷ See, e.g., Gallow (2022) for an overview.

and determinism,²⁸ and it may well seem that a focus on top-down causation will not be a good way to resolve the issue to the satisfaction of a relevant portion of the parties involved.

2.2.3.2 Emergence

This leaves us with the term “emergence”, which at least has the benefit of being a neologism open to a clear technical definition without much interference from everyday usage. The term originates in biology, in the late nineteenth century discussion about a possible middle path between a reductively materialistic understanding of life and the positing of extraphysical vital principles, but aspects of the philosophical position can be traced back to antiquity.²⁹ That historical lineage with its emphasis on the example of life may, however, also be somewhat problematic. Starting with the example of life would *guarantee* the existence of a witness—it is uncontroversial that life is real. The technical details of the definition of emergence could then, as it were, be treated as free parameters to be adjusted to fit that witnessing example, which might look question-begging. It would be much better to start from an independently motivated understanding of emergence that connects that notion with our question about ontological levels, and *then* to see whether we can find plausible instances.

Looking at the literature, one finds a fair number of definitions of emergence, all of which of course have some relevant motivation behind them. We will not attempt to give a systematic overview.³⁰ One distinction is, however, crucial: On some definitions, emergence turns out to be an *epistemic* notion, while on others it marks an *ontological* distinction. It appears to be unquestionable that there are witnesses of even quite strict definitions of epistemic emergence, for which the leading idea is that higher-level properties cannot be deduced from the full set of lower-level properties even in principle. On many definitions, any case of quantum entanglement (a phenomenon that we will discuss starting in Sect. 3.4.3) already witnesses epistemic emergence, as one cannot compute the state of an entangled joint system from the properties (reduced states) of its parts.³¹

²⁸ If causation has to be deterministic, as many hold (meaning that the cause *necessitates* its effect), then it seems that higher-level causes must be mere epiphenomena, as the lower level causes already force the outcome at all scales. See the references in footnote 26. By way of contrast, see Anscombe (1971) for a classic defence of indeterministic causation, and see the papers in the collection by Mulder et al. (2022) for recent discussions of Anscombe’s views.

²⁹ See O’Connor (2021) for an overview, including a reference to Lewes (1875), who coined the term. Meincke (2024) highlights differences between the Aristotelian tradition and the British emergentists.

³⁰ Recent contributions include, e.g., Yates (2016) and Humphreys (2016). See also the collection by Sartenaer and Baysan (2021) and footnotes 29, 31, and 32.

³¹ See Howard (2007). See also Butterfield (2011a,b) for detailed discussion of this and further cases.

Arguably, an even stronger case in point is the resultant total magnetisation of a ferromagnet given an initial configuration of the interacting spins.³² Epistemic emergence may be surprising at first sight, but it does not yet properly address the problem of ontological levels. Therefore, we are here only concerned with ontological emergence, which is the stronger and more controversial notion.

2.2.3.3 Ontological Emergence

Ontological emergence means that with respect to a lower level that is accepted as *really* existing, at a higher, emergent level of organisation there exists something really novel, something not just surprising to us (which would indicate epistemic emergence), but constituting an independent and in some sense autonomous reality. The guiding idea is, thus, exactly the same as that behind the notion of higher ontological levels that we have been using so far. Connecting these two notions is useful because in philosophy of science, there is by now a fairly substantial literature on specific purported examples of ontological emergence within physics. For our purposes, these discussions are more helpful than the large literature on emergence in the philosophy of mind: Unlike the prime example in that domain, consciousness, the physical phenomena in question (such as superconductivity or critical opalescence) are not themselves philosophically contentious.

In discussions of ontological emergence in physics, one thus considers cases of empirically well-established macro-phenomena that (as everybody agrees) arise in systems of (very many) well-understood micro-constituents. In such cases, the question of what ontological novelty could mean becomes especially pressing. The title of the classic paper by Anderson (1972), “More is different”, already points to the fact that novelty must in some sense be connected to *qualitative* differences arising out of *quantitative* ones. This is an idea that is very well known and practically useful (a short string of wool doesn’t keep you warm, but a woollen sweater does), but its philosophical soundness is contested, perhaps also due to its ancestry.³³ This problem is aggravated by the fact that in many physical examples, models for qualitatively different system behaviour (e.g., for symmetry breaking or for phase transitions) force one to consider infinite systems (the thermodynamic limit of $N \rightarrow \infty$), while empirically, the qualitative difference of a novel phenomenon is of course found in real, finite systems, and the physicality of infinite models appears questionable.³⁴

³² Humphreys (2015) critically discusses a relevant undecidability result of Gu et al. (2009).

³³ See, e.g., Carneiro (2000) for an overview that traces the use of the idea of “quality from quantity” from Hegel via Marx and Engels to modern applications. Anderson (1972) is aware of that ancestry, as he makes clear at the end of his paper.

³⁴ See, e.g., the papers in Falkenburg and Morrison (2015) for a number of case studies highlighting these difficulties. See also Landsman (2017, Ch. 10) on what he calls Earman’s and Butterfield’s principles, which demand, respectively, that real physical effects must not depend on idealisations (ibid., p. 367) and that, more specifically, robust novel behaviour must be present for finite systems, before taking the limit of $N \rightarrow \infty$ (ibid., p. 370).

In our view, what is missing in most of these debates is a focus on the positive role that indeterminism plays in enabling emergence. In fact, we will argue next that unless a lower level affords enough open possibilities that can be organised at a higher level, that higher level cannot be real.

2.2.3.4 Defining Ontologically Emergent Individuals

While we agree that there is much to be learned from the study of phase transitions, symmetry breaking, and the thermodynamic limit, we propose a fresh approach to emergence, based on indeterminism and focusing on concrete individual entities. Our general picture of ontological emergence is this. There has to be a lower level of matter whose interaction conforms to laws that are indeterministic, i.e., laws that, while constraining what can happen, do not normally prescribe a unique later configuration based on a given current configuration.³⁵ And constituted out of that matter, there has to be a concrete entity with surprising, autonomous features, which we describe as a higher level of organisation, thus witnessing some form of epistemic emergence (which is *not* yet our target). That higher-level entity will be an *ontologically emergent individual* *via-à-vis* the lower level if it independently organises the lower-level matter as material.³⁶ This organising must then of course be empirically manifest in processes by which the individual in some sense makes use of the lower-level indeterminism, e.g., by leading to the realisation of specific lower-level possibilities in a dynamic equilibrium, or by influencing the statistics of lower-level random happenings.³⁷ The benefit of this approach to emergence is, first, that it targets concrete individuals such as individual agents, so that it offers a chance of helping to pass a verdict on the question that interests us here, namely, *whether there really are agents*. And second, it brings into view two aspects that other approaches often leave out, namely, the concrete individual's

³⁵ From the point of view of the lower level, that is, the future is open. The laws governing that level are therefore, to use an analogy offered by Anscombe (1971, 143), like the rules of chess rather than like the rules of an infantile card game: “the play is seldom determined, though nobody breaks the rules”.

³⁶ See Mulder (2016, 2021) for background on the idea of “matter as material”. In the former paper, Mulder briefly discusses the notions of emergence and downward causation, but rejects them as unhelpful for his specific aim of providing “a vital challenge for materialism”. In the latter paper, he avoids the terms “emergence”, “top-down causation”, and “holism” altogether. We here take the liberty of connecting his thoughts with a discussion that targets emergence head-on.

³⁷ We will come back to this topic and discuss the apparent tension with lower-level indeterminism in Sect. 3.6.2. For reasons given above, we avoid the term “top-down causation” here, especially as that often comes with deterministic connotations. It is crucial for our view that the lower level afford real open possibilities, thus providing wiggle room that is needed for any higher-level organising to have an effect at all.—We can leave it open whether the emergent higher-level entities are understood as substances or as processes, i.e., whether to subscribe to a substance- or to a process-ontology. The former has tradition on its side and connects well with our *de facto* metaphysics (Strawson 1959), but the latter has much going for it, especially concerning biological entities (Nicholson & Dupré 2018).

history (how it originates and what its possible courses of development are) and, in connection with that, the *environment* or *niche* in which that individual could originate and can develop. In a nutshell, thus, an ontologically emergent individual is an entity, (1) consisting of matter that provides indeterministic open possibilities as a resource, that (2) organises that matter as its material and that (3) exists in a specific environment that allows for this specific kind of organisation.

To illustrate these thoughts via a preliminary example that is still too simple, consider a house. Are there houses, really? On an ontologically minimalist view, a house is just “matter arranged house-wise”. And indeed, a house consists of bricks, beams, mortar, and so on, suitably arranged. The matter that is so combined is here used as material in the literal sense, as building material. That material can be so combined because it affords a large enough range of possibilities for combination. When the individual house has been built, out of these many possibilities, one has been realised. In this case, it will have been realised by human agents at the service of their needs and interests, following a process of building that starts with laying foundations and then adding walls and floors and a roof, in an environment in which houses can persist (not inside a volcano or on an ice floe, for example). Now with the house, new higher-order properties and phenomena come up that are not simple aggregates of the properties of the materials used. For example, the house has a specific enclosed volume—but the materials used have a much smaller volume, and they could also have been used to build a house that encloses a smaller or a larger volume. The house also has a street address, which makes no sense for the materials. The house has a solid or a leaky roof, as the case may be. And the house may be beautiful even though the materials used to build it were not, or the other way round.

That may be good enough for everyday purposes, but as a witness of individual ontological emergence, the example of the house will probably not convince the sceptic, who may argue that the example presupposes emergent entities at a similar or even at a higher ontological level to begin with. The bricks and beams that furnish the underlying matter for the organisation of the house are themselves ontologically individuated at a fairly high level, and using them in the example might smuggle in the level of emergence that is then claimed for the house. Worse still, in describing the example of the house we had to have recourse to agents and their interests, because houses (in fact, already bricks and beams) are artefacts. Again, this may be criticised as presupposing emergent higher levels to begin with. In order to illustrate the notion of individual ontological emergence without such extra assumptions, we propose to look at a different, thoroughly physical example: the sun. We want to defend the thesis that *the sun is real*. Really.

2.2.3.5 The Sun Is More Than its Parts

The sun has been around long before us. It is one of about 10^{11} stars in our galaxy, the Milky Way, which is one of the 10^{11} galaxies in the universe.³⁸ A lot is known

³⁸ These numbers are lower bounds based on current means of observation; see Marov (2015, 280).

about the dynamics and the development of stars in general, and a lot more is known specifically about the sun—its origins and history, its chemical constitution and the nuclear reactions going on in its core, the way it produces visible light and other radiation sustaining life on earth, and much else including its eventual fate when its nuclear fuel will burn out in a couple of billion years.

The sun was personified as a god in many ancient religions. Given our present understanding of the heavens, that is certainly a classic example of superstition rife for eliminative reduction: the sun-god isn't real. But what about the sun, our familiar star? There *is* something there at the centre of our solar system, for sure—but can we defend the ontological reality of the sun against the non-eliminative reductionist charge that all there is out there is (a lot of) “matter arranged star-wise”? Can we defend the sun as an exemplary ontologically emergent individual? We believe that the sun is a good example for following the dialectics of that issue exactly, and we will defend a positive answer: the sun is really real.

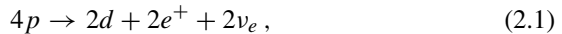
In a way, of course, nobody denies that. But then, nobody denies that there are cats, and still the non-eliminative reductionist about life will hold that the notion of a cat is just a useful crutch for us, but not descriptive of ontological reality. In the reductionist's reality, there is only matter, and in the case of Tibbles the cat, that matter is arranged cat-wise. This position, of course, threatens the independent reality not just of Tibbles the cat, but also of Tibbles the feline agent—and of all other material agents along with him in one fell swoop. If we can defend the example of the sun as an ontologically emergent individual, we will have established a counterinstance to the general reductionist claim that “*X*'s are nothing but matter arranged *X*-wise”, and that will help save Tibbles, and agency, in the end. The good thing about the example of the sun is that there will be no need to refer to biology, life-processes, teleology, or other notions that already belong to a higher level of organisation. We have a generally accepted, convincing account of how the sun came to be and what sustains its current form and operation, all based on fundamental physics and without the need for a long story of (Darwinian) evolution. That account of what the sun is, in our view, provides a strong argument in favour of its ontologically emergent reality.

The history of the sun is, very roughly, as follows. According to the currently accepted Λ CDM-model of the history of the universe (“ Λ ” for a positive cosmological constant, and “CDM” for “cold dark matter”), the big bang ca. 13.8 billion years ago was followed by a swift, vast expansion of the universe. During an early phase of primordial nucleosynthesis, the lightest chemical elements hydrogen and helium, and trace amounts of other light elements, were formed and distributed almost, but not quite, uniformly as dust throughout the expanding universe. In some places the dust gathered through gravitational attraction, leading to the formation of the first generation of stars. Through nuclear reactions in these stars, supernova-like explosions and further processes, other chemical elements were also formed, and thus was produced the matter that constitutes stars like the sun, and which is used as material in them. Around five billion years ago, the sun and the solar system formed via the aggregation and subsequent gravitational collapse of a cloud of dust,

and through the action of gravity, the core of the sun became so dense and hot that collisions of hydrogen nuclei there regularly lead to nuclear fusion.

The understanding of the processes by which the sun and other stars produce the energy they radiate is one of the triumphs of quantum physics. In stars, mass is converted to energy, illustrating Einstein's famous formula $E = mc^2$, at an astonishing rate; the sun has converted roughly 100 earth masses to energy so far. The basic processes are fairly simple, and their quantitative features can be computed with high precision. At the core of these processes are electroweak interactions through which, via the change of an up quark to a down quark, a (positively charged) proton is converted to a neutron, a positron (the positively charged antiparticle to the electron), and a neutrino. In what follows, it will be important to note that nuclear fusion must be described by indeterministic, relativistic (quantum field theoretic) models, since particle numbers are not preserved in these reactions.

At the level of atomic nuclei, there are several paths by which hydrogen nuclei (protons) fuse to produce the very stable nuclei of helium-4, which are currently the end products of the sun's energy production. (They will be used up at a later, helium-burning stage to form carbon.) One main path of hydrogen fusion consists of the following "p-p" (proton-proton) chain reaction:



Overall, this chain of reactions uses up four protons (p), or hydrogen nuclei, to produce one helium-4 nucleus (^4He ; consisting, like the α particles resulting from familiar radioactive α -decay, of two neutrons and two protons), two gamma-ray photons (γ), two electron neutrinos (ν_e), and two positrons (e^+), which are the antiparticles of electrons and which are subsequently annihilated by collisions with ambient electrons, producing more gammas. As intermediate products, hydrogen-2 nuclei (deuterium, d) and helium-3 nuclei (^3He) are produced, but then used up. The total energy that is released in one such chain is 26.7 MeV. Apart from the main branch of the p-p-chain (eqs. 2.1–2.3), which accounts for 69% of reactions involving ^3He , there are two more p-p chains, both of which start with



From there, the main branch (99.7%) continues via ^7Li and p to produce two nuclei of helium-4, while another branch (the remaining 0.3%, i.e., a total of around 0.1% of all hydrogen fusion) produces ^8B as an intermediate product, which then decays to also form two helium-4 nuclei. (Other, more complex reactions also contribute to the sun's energy production to the extent of around 1%.) Theoretical calculations of these ratios are based on fundamentally indeterministic models of the scattering processes involved, and they are supported, e.g., by empirical studies

of the spectrum of solar neutrinos.³⁹ We have come a long way from the sun-god Helios driving a chariot through the sky—we pretty much know what makes the sun shine.

But given that we know so much about the processes that go on at the centre of our solar system, aren't we facing a "disappearing sun worry" analogous to the disappearing agent worry mentioned above? Why should we posit the sun as a real individual entity rather than acknowledge that it is just a part of the overall happenings in nature—a part that we humans understandably, but still (from an ontological point of view) arbitrarily focus on?

On the basis of our definition of an ontologically emergent individual given above, we can argue that our focus is not arbitrary, and that the sun is real. In order to exist as an ontologically emergent individual, the sun requires (1) a lower level of matter that provides indeterminism as a resource and (2) the higher level of the individual that organises or uses that matter as material. We also said that (3) real individuals normally exist in specific environments, or niches. To provide our argument for the reality of the sun, we can show that all these three points are fulfilled.

As to (1), lower-level indeterminism: all the processes involved in nuclear fusion are fundamentally quantum-mechanical, and therefore indeterministic. But of course, saying this is controversial given debates about quantum foundations. We will have some more to say about this issue in Chap. 3. In the case of the sun, however, there are some specific details that strengthen the case for indeterminism. Consider the start of the p-p chain, Eq. (2.1), which describes the production of one deuterium (hydrogen-2) nucleus from two hydrogen-1 nuclei (protons). That process is exothermic, releasing energy, but it requires an environment like that maintained in the sun's core to function. Specifically, it requires the two protons to get together to within the range of the weak force, through which one of the up quarks of one proton can (with a rather low probability) be converted to a down quark, a positron and a neutrino, and the resulting proton and neutron can then be bound by the strong nuclear force to form a deuterium nucleus. At the very short distances involved (less than 10^{-17} m), a huge electromagnetic repulsion needs to be overcome, and very few proton-proton interactions will have enough energy for that. The process is therefore mainly driven by indeterministic quantum-mechanical tunnelling: even allowing for a classical picture of particles with definite momenta moving on localised trajectories, one can only compute probabilities for the protons' getting within the range of the weak force. Even more importantly, the weak interaction involved then changes the particle numbers. In contradistinction to non-relativistic quantum mechanics, no deterministic models for these interaction have been worked out in detail, and it appears implausible that such models can

³⁹ See, e.g., LeBlanc (2010, Ch. 6) for a textbook overview of the theory, and see the Borexino Collaboration (2018) for recent experimental results.

be established.⁴⁰ All models for what goes on in the sun at the level of matter are fundamentally indeterministic.

As to (2), the sun indeed makes use of its matter as material, in the sense that the matter undergoing fusion processes is organised to keep the sun in a dynamical equilibrium. The radiation pressure from the fusion processes exactly balances the pressure exerted through gravitational attraction. Of course, there is no agency or intentionality involved—the processes organise themselves. But they organise themselves in a way that the temperature in the sun’s core is maintained such that exactly the right amount of fusion takes place that keeps the sun in a stable dynamic equilibrium. Over the course of its history, the sun has had different conditions in its core—both its chemical composition and its mass and, accordingly, its core temperature have been and are changing slowly, but constantly. A dynamic equilibrium has been and still can be maintained because the sun’s matter affords the required possibilities for adaptation. In stark contrast to a static equilibrium, maintaining a dynamic equilibrium requires flexibility of the underlying material to allow for such constant adaptation of conditions. In case the required flexibility is no longer given (e.g., when the sun will have burnt too much of its fuel to maintain the required radiation pressure through fusion), the organisation will break down, and the sun will collapse, which will mark the end of its existence as the star we know. Thus, in contradistinction to a mere aggregate of matter, we have here underlying matter that affords possibilities for its organisation, and a corresponding organisation at a higher level—the sun—that regulates the conditions under which that matter realises its possibilities, and in such a way that the sun keeps shining.

Regarding condition (3), the ecological niche, it pays off to look at the reductionistic doctrine of a mere aggregate, and in particular at the slogan “a star is matter arranged star-wise”, in more detail. An aggregate is present whenever its parts are present in the required arrangement, so by the slogan, a star should be present whenever matter arranged star-wise is present. But this is demonstrably

⁴⁰ For non-relativistic quantum mechanics, Bohmian mechanics is often mentioned as an important deterministic alternative to the indeterministic view of quantum mechanics that we will defend in Chap. 3. There are treatments of tunnelling in the Bohmian picture (see, e.g., Douguet and Bartschat 2018), but these are viewed critically (see, e.g., Hofmann et al. 2021). Importantly, for the relativistic case involving particle creation and annihilation that is important here and for interactions involving Fermions (such as protons, neutrons and electrons), i.e., for what goes on in the sun, the literature on Bohmian mechanics only contains a sketch of an approach, which is itself fundamentally indeterministic, and this for reasons of principle (Dürr et al. 2004). This assessment is corroborated by a recent overview article by one of the main proponents of the approach (Tumulka 2021); see also Maudlin (2019, 223). Perhaps similarly, current expositions of the “many worlds” interpretation of quantum mechanics, which is also claimed to be deterministic, start with a “wave function of the universe” that assumes a fixed number N of particles (Vaidman 2021, §6.3), and may therefore also be in need of an indeterministic supplement to treat the present case. (At least we are not aware of an explicit, detailed discussion of the splitting of worlds in particle creation or annihilation.) This leaves superdeterministic approaches such as t’Hooft (2016), which however involve a strong form of agency scepticism, which we discuss in Sect. 2.4.2. See also note 148 to Chap. 3.

false. Suppose that matter arranged exactly like the sun's matter right now was present as a (hugely improbable) fluctuation in the very early universe. Then that fluctuation would be wiped out in a matter of fractions of a second. It would not be true to say that processes of nuclear fusion were going on—at those temperatures, all there was a quark-gluon plasma. Not even atomic nuclei were present back then, even if “matter arranged nucleon-wise” was present. This is a consequence of the commonly accepted history of big bang nucleosynthesis. For example, the existence of the first nuclei of deuterium (d) depends on the freezing out of the equilibrium of creation and destruction in the reaction $p + n \rightarrow d + \gamma$, for which temperatures must have fallen to 0.1 MeV (Cyburt et al. 2016, §II.A). This does not mean that the reaction cannot happen at higher temperatures. But if it happens, reactions in the other direction that break up the resulting compound also happen, and so there is no stable binding of deuterium nuclei; the “matter arranged deuterium-wise” is, at such temperatures, a mere fluctuation, an accidental happening, and does not constitute atomic nuclei. One relevant time scale for the assessment of stability of deuterium can be derived from comparing it to the radioactive isotope of hydrogen, tritium (^3H), which has a half live of just over 12 years. For a mere fluctuation with a life-time much smaller than this, the essential properties of stability (in deuterium) vs. radioactivity (in tritium) make no sense, and this is another indication that in the very early universe, matter arranged deuterium-wise would not have been deuterium.

The point of these thought experiments is this: For a star to be a star, it has to be present in an environment in which atomic nuclei are generally stable, and specifically available for fusion, and this presupposes a certain temperature range and a certain stage of cosmogenesis. *The exact same arrangement of matter* would *not* be a shining star in a different, much hotter environment, in which it could only arise as a chance fluctuation.⁴¹ Stars do not come about in that way; they have a fairly well understood history, as sketched above. There is more to a star than its current instantaneous material composition: there is also its history and, if you allow, its (quasi-ecological) niche. Given this nexus of dependencies, we are justified in taking stars to be ontologically emergent real individuals rather than aggregates of matter. The sun is more than its parts.

2.2.3.6 The Immune System Is Real

We treated the case of the sun as an ontologically emergent individual at some length because that case exemplifies a purely physical system that still exhibits the “matter

⁴¹ In a move that is characteristic of variants of global scepticism, the reductionist can at that point enlarge the system and include a part of the sun's actual environment, e.g., our solar system. This will, however, not help—we can repeat the chance fluctuation thought experiment with *that* system as well, showing that what was identified was not a star plus some of its environment, but again just a cosmic fluke that just lasts a little longer. In the end, the only stable position for the reductionist will be to include the whole universe, completely losing track of the sun. The conclusion is, at maximum, the tautology that the universe with the sun in it is equal to the universe with the sun in it—but this will not provide any comfort to the reductionist any more.

as material” structure that characterises higher ontological levels. In living systems, there are of course even more obvious cases of systems that make use of the open possibilities provided by their matter’s underlying indeterminism. A particularly striking example is the process of somatic hypermutation, which is behind the highly adaptive reaction of the immune system of vertebrates to a wide range of pathogens. Briefly, the B-cells in the immune system can produce specific immunoglobulins that bind to a pathogen to neutralise it. As the range of possible pathogens is more or less unlimited, specific immunoglobulins that are effective against a new pathogen cannot be present in reserve. They can, however, be produced in an adaptive process that implements the evolutionary recipe of mutation and selection. There are specific sites on the DNA coding for immunoglobulins for which the rate of mutations is a million times higher than for other sites. The molecular mechanisms involved in this so-called somatic hypermutation are still being studied intensively (Odegard & Schatz 2006; Chi et al. 2020). When an immunoglobulin that has come about through this random process has proven effective against a currently present antibody generator (antigen), its production is enhanced so that the specific immune response is strengthened. Among other benefits, this mechanism is also responsible for the effectiveness of vaccines, as it leads to a long-term memory effect.

In this process, we see in an exemplary way how a higher-level entity (the immune system) makes use of lower-level indeterminism. In this case, it is exactly the random realisation of open possibilities that drives the mutations and thereby provides the material (the different variants of immunoglobulins) that is then used (differentially copied or suppressed) through the higher-order system. The immune system is part of the life processes going on in a vertebrate individual, constituting one of its organised sub-units (organs).⁴² The process is a clear example of the use of indeterminism as a resource, or, in the terminology of Noble and Noble (2021), the harnessing of stochasticity. As we have argued, such uses of indeterminism are at the root of ontological emergence.

2.2.3.7 From the Sun and the Immune System to Agency

On our view of emergent individuals, ontological categories arise through hierarchies of “matter used as material”, which presupposes that the matter in question afford a wide enough range of possibilities for combination so that it can be used as material. This can be true or false, and we can at least try to find out which. It follows that the real existence of separate, ontologically respectable levels of organisation in nature is something that science can study and explain. Life organises biochemical matter to sustain the various homeostatic and homeorhetic processes that make up a living organism (Dupré 2021), and this is made possible by the fact that the respective matter affords enough possibilities for combination. Birds can build nests

⁴² This organ is distributed over the whole body, with the bone marrow playing a key role (hence the “B” in “B-cell”). In terms of mass of the organ, in humans the immune system is comparable to the brain (Alberts et al. 2015, 1308).

from twigs, and people can build houses from bricks and beams, because these materials afford enough relevant possibilities for combination. Two people can carry a table across a room as a real joint social effort because they both have the relevant skills and the possibilities to combine their individual efforts.

The positive examples of ontological emergence that we have spelled out affirm the possibility of the widespread occurrence of higher-level entities, including agency. We have thereby transformed the dialectical situation sketched at the beginning of Sect. 2.2.3: The reality of agency is not just a hypothesis that has not yet been successfully refuted by the reductionist—it is a claim that can be positively worked out and defended by providing details along the lines of our examples. In this way, we can embark on a journey to show that and how agency exists as a stable higher level of organisation, something that life on earth provides room for and which is really distinct as an ontological category. The challenge is to work out a detailed account of agency via the method of explication, and then to show how agents can organise their underlying matter that offers the relevant possibilities—i.e., their bodies—in such a way that organisation is properly the agents'. If that picture can be sustained, then it is not only a fact that we use the category of agency to make distinctions in the world that are important for us—we will be able to show that these distinctions also have a real basis in what the world is like.

The next steps on our journey, before describing any concrete modelling efforts, must be to delineate our target notion of agency in more detail. Even if our interest is primarily ontological, we still have to provide a more detailed explication of the notion of agency, which starts from our practices involving that notion.

2.3 Approaching the Agency Question by Way of Our Practices

As in the case of numbers, the specific question of agency as an ontological category has to be well separated from simpler, ground-level scientific questions of classification and existence. A positive answer to the question whether 7 is a prime number does not settle the ontological question whether numbers exist. Similarly, a scientific answer to the question whether the earthworm is an agent does not settle the ontological question whether agents exist. The answer to the ground-level scientific question about the agency of the earthworm surely depends on the exact contours of the notion of agency that we are in the process of explicating, but the method to answer it will be a scientific, biological and not a philosophical one. The ontological question about agency, on the other hand, operates at a higher level of abstraction: Are there agents, *in addition to our practices of taking certain things (e.g., cats, or each other) to be agents*, of attributing agency to certain material things and of viewing them with something like an “intentional stance” (Dennett 1984)? Here we may still feel the pull of Nagel’s worry: is it really necessary to posit agents, given that “[e]verything I do is part of something I don’t do, because I am a part of the world” (Nagel 1986, 114)?

In the previous section we have provided a template for ontologically differentiating among the “parts of the world” in a non-arbitrary fashion. But for agency, the issue is especially delicate because of the autonomy of agents and because of their special influence on what happens in the world. In addition, questions about agency are complicated by unclarity about what agency is, which relates to the relative scarcity of discussions of the notion of agency in the philosophical literature. As we already said, this means that it will be a large part of our work here to get a good fix on what agency is. Such a clarification—an *explication of the notion of agency*—is not only needed for ontological purposes, but also for answering the scientific question of whether certain animals are agents or not. Which standards do we use to decide? The scientific and the philosophical questions are interdependent, and *prima facie*, there are several conflicting ways of addressing them.

In philosophy, agency is mostly discussed with reference to human beings, which means that the agents in question are very special in that they are also linguistic and moral beings. Even if morality is officially left out of the picture, agency is often connected with full-blown intentional states, and according to some influential lines of thinking (e.g., Davidson 1978), this presupposes that agents pass propositional judgements and have a language—which may introduce morality through the back door: One can argue that an agent who has a language thereby has available a medium for communication that also allows for reflective thought about how its actions will impact others, and that capacity may already ground a duty to exercise it. In contrast to this high-level notion of agency in philosophy, in computer science, a very thin notion of agency has been found useful—an agent in that sense can be any system that interacts with its environment through sensors and actuators (Russell & Norvig 2022). So, part of our work will be to find a way in between these two extremes in order to home in on a notion of agency that is both ontologically well-grounded and clear enough to be a target for modelling.

Life as a Separate Ontological Category? We can illustrate the challenge by comparison with the situation with respect to life. Like for agency, the question whether we need a separate ontological category of living beings is controversial. But that question has a long and rich history, both in philosophy and in science.⁴³ Furthermore, it operates before the background of a well-understood and largely uncontroversial set of instances of living beings. We know pretty much what we are talking about here: algae and trees, mushrooms and jellyfish, mice and men. There may be borderline cases, like viruses, and the exact criteria for life vs. death are somewhat contentious. There is also the controversial question of the possibility of artificial life. But the question of whether life should be treated as

⁴³ Life as a distinct order of being, with its own method of investigation, is treated, e.g., in Aristotle’s *De partibus animalium*, and Hegel (1812, 761) envisaged a “logic of life”. See also Driesch (1914) for the controversial topic of entelechy.—As outlined above in Sect. 2.2.2, the status of life and the relation of biology to physics play a key role in debates about reductionism in philosophy of science. Meanwhile, philosophy of biology has been established as a separate sub-discipline. For relevant further references, see Mulder (2016).

a separate ontological category is not relevantly affected by these borderline issues. As outlined above, reductionists, who hold that a separate ontological category is superfluous, simply point to the fact that all of life is constituted by complex chemical processes, and that this holds across the board. Vitalist ideas of a specific immaterial energy or force attached to life (called “entelechy” by Driesch 1914) were abandoned in the early twentieth century, and long before that, the idea of still-so-called organic chemistry as a special realm of matter had already been rejected. Physically and chemically, life is of a piece with the rest of what goes on in the world, it is just materially specific (e.g., homeostatic and carbon-based) and probably more complex than anything else in the world. The necessity of an ontological category of the living has to be argued for before that background. If life is special as a kind of being, that claim needs to be established without special pleading, before the background of a physically and chemically unified world. This also means that if a reductive account of life prevails in the end, it will have to be of the non-eliminative type, proclaiming life to be real, but “nothing but” physics and chemistry packaged up in complex ways, as *biophysics* and *biochemistry*, as one will say. These matters are complicated and subject to much controversy. But their focus is at least clear and uncontroversial—in contrast to the ontological question about agency, the ontological question about life targets a well delineated and extremely well studied class of entities.

Agency as a Separate Ontological Category? As we already mentioned, in contrast to the case of life, the question of agency as a separate ontological category is not much studied. We know of agency first hand, and agency is supremely important for us—but the prime examples of agency, we ourselves, are also moral agents, and many philosophical discussions therefore run together agency in the sense that interests us here, which is exhibited widely across the animal kingdom, with moral agency, which is (arguably) only exemplified by us human beings. So, there is a lack of separation of agency upwards on the ontological scale. A similar problem concerns the separation downwards: all agents we currently know of are alive. Does this mean that agency is a specific form of life, or is agency a notion that can also be exhibited by technological artefacts, which are not alive? This issue is pressing in our context, as we are interested in modelling agency in a way that could underwrite the possibility of artificial agency, perhaps even indicating a path towards its realisation. That path might be blocked if agency was necessarily tied to life. Now it appears that agency and life are fairly close indeed: both can certainly only be instantiated by rather complex causal processes. But agency and life also appear to be orthogonal in that life is, at least according to some accounts, tied to the historical order of living beings on our planet,⁴⁴ while agency is a more abstract characteristic of autonomous entities that may be independent of certain necessary features of the living. For example, it seems that reproduction, some form of copying, is characteristic of life, but it is at least not obvious that this is also the

⁴⁴ This is argued for by Thompson (2008, 39ff.).

case for agency. In connection with this observation, it also appears pertinent that current developments in Artificial Intelligence do target artificial agency,⁴⁵ while the topic of artificial life appears to be more of a niche issue that is rather separate from core developments in AI.

The comparison with the case of life thus confirms that a discussion of the ontological status of agency should be prepared via a thorough explication of what agency *is*. For this it seems best to start from our everyday practices. We will focus on two such practices concerning agency: how we talk about agency (Sect. 2.3.1) and how we attribute it (Sect. 2.3.2).

2.3.1 Agency Talk

In philosophy, it is often helpful to take a close look at the way in which a phenomenon in question shows up in our language.⁴⁶ Of course, the way we talk does not settle ontological matters once and for all. Language is a product of biological and cultural evolution, and the distinctions it incorporates are historically contingent. But the way we handle ontological categories in language does tell us something at least about our cognitive organisation, about what we *take to be* relevant categories of being. And as we have evolved in the real world, it is plausible to assume that the organisation of the real world has somehow left its mark here. The fact that we use the category of agency widely certainly has repercussions for how we talk, and this in turn provides support for an ontologically pluralist view that admits agency as a separate ontological level.

Basic Agency Talk Here are some initial relevant observations. When we talk about agency, we talk about an ability or a capacity of certain entities, agents. An agent *does* things, in a sense in which non-agents don't *do* things. This is true and fundamental, but not yet very helpful. How can we get a fix on the special type of *doing something* that is involved in agency? Linguistically, what an agent does is reported in the *active voice*. Agents occupy the subject position in sentences involving transitive causal verbs. Peter cut the cake. Jill is feeding the birds. The cat will catch a mouse.

Ability to function as a subject in activity reports is indeed a necessary condition for being an agent, for possessing agency—but it is not sufficient. The stone broke the window, the cake delighted Paul, the square root of -1 bewildered Törleß. Almost anything can function as a subject in such sentences. But with agents, there is a special sense of involvement in the activity, one that is absent from non-agentive

⁴⁵ See, e.g., Wu and Tegmark (2019), Fossa (2021), and Thomason and Horty (2022).

⁴⁶ One could say that the linguistic turn of the mid-twentieth century went overboard with this idea and for a while tried to account for all matters of philosophy via facts about language. It is now generally acknowledged that cannot be all. But considerations of language are often highly important nonetheless, as we will indicate.

reports. Circularity threatens, of course: We cannot say that the involvement of the agent is an active one, and leave it at that. That is true, but it would not be informative. Can circularity be avoided? Certainly we also need to look at other aspects of agency talk, starting with the specific type of predication involved.

Temporal vs. Atemporal Predication A very basic ontological distinction, that between abstract and concrete entities, already shows at the level of one of the simplest ways of building sentences, predication. The basic form of predication, *Fa* (“*a* is *F*”), applies already to abstract entities—witness “7 is a prime number”. For concrete entities, however, predication is temporal, expressing the fact that *a* is *F* *now*. This may have been different yesterday, or change at some point in the future. Tibbles the cat is hungry now but he wasn’t hungry five hours ago, and let us hope that his owner remembered to get him cat food at the supermarket, so that he won’t be hungry much longer. Such temporal distinctions do not make sense for abstract entities, and this shows in language: The sentence “7 was a prime number in 1853” is odd, to say the least.⁴⁷ Change, the possibility of acquiring new properties, only makes sense for concrete entities. But it makes sense for all concrete material entities, no matter their organisational complexity, and thus it does not yet distinguish agents from non-agents.

Needs and Wants Further distinctions, which are specific to agency, are more subtle. For example, we can ascribe *needs* to certain concrete beings, but not to others. Living beings clearly need things—food, shelter, air to breathe or whatever the case may be. This is specific for the ontological category of living beings. Rocks don’t need anything.⁴⁸

But what about artefacts? Doesn’t a computer need electricity in the same way in which a cat needs water? This may sound suggestive, and we understand what we mean when we talk that way, but the analogy is nonetheless faulty. The live cat needs water, full stop. The computer needs electricity *in order to function as our word processor*. If we use it in order to smash in a window or to block our car from rolling down the hill, the computer does *not* need electricity. The need thus applies to the computer only courtesy of our own goals and interests. *We* may need the computer to have electricity, but the computer by itself needs nothing.

⁴⁷ A context in which we would use a sentence like this would be in indicating that a definition has changed, and so the sentence would in the end be about our historical practices and not about the abstract entities involved. Thus, one might say that “in the early nineteenth century, all continuous functions were differentiable”, meaning that the definition of continuity changed at around that time. We understand “Pluto was a planet in 1960” along similar lines.

⁴⁸ In line with our considerations of the importance of niches for the existence of entities in Sect. 2.2.3 above, we can say that rocks can only exist at temperatures that are low enough for solid matter to persist. In this sense, rocks need sufficiently cool temperatures. The difference vis-à-vis the needs of living beings is that if the needed conditions are not given, rocks simply cease to exist, so that these needs are, really, necessary conditions of existence. For living beings, on the other hand, the frustration of their needs is primarily just *bad* for them. Of course, unfulfilled needs may eventually lead to death, depending on which need is at issue.

What about wants?⁴⁹ Again, we may say that the computer wants user input at a certain stage in processing, but again this is figurative speech that depends on what *we* ourselves want. And we *do* want things.

Wants indeed seem to be specific for agency. Not all living beings can be said to *want* things. Algae, bacteria and cacti surely *need* certain environmental conditions, but they do not *want* anything specifically. On the other hand, cats do want certain things—e.g., Tibbles the cat may currently want his owner’s attention—and not others—e.g., he does not want to be sprayed with water. Literal ascription of wants appears to single out an entity as an agent. Entities that want something have acquired specific aims or goals that pertain to them as individual agents with their own history, and the fulfilment of these wants makes a difference to them.

Imperatives Here is another pertinent linguistic fact. It is obvious that all users of imperatives are agents, because issuing an imperative is a speech act, and only linguistic beings, that is, we ourselves (who *are* agents), can do things with words. But more interestingly, the addressees of imperatives also need to be agents. Imperatives can only be directed at agents. And the class of addressees here is larger than just us: We can utter imperatives to dogs and horses, given that these are trained to respond to them, and you might even try addressing Tibbles the cat in that way. Of course, figurative uses of imperatives are common—we speak of imperative programming languages, for example, and we can issue commands to digital assistants, to robots, and to other artefacts. But as with a motor’s need for oil, our way of talking here is strictly dependent on ourselves and our practices. On the other hand, it is interesting that some living beings *can* be addressed with imperatives in much the same way that we can. The common ground, it seems, is that we and some animals *who are also agents* can be trained to respond to imperatives. In fact, the issue of training and learning will play a crucially important role in our discussion of characteristics of agency in Sect. 2.3.3 below.

A Further Aspect of Agency Talk? The Imperfective Paradox We mention briefly, only to leave it to the side, a further linguistic observation about the reporting of events that may be helpful, but which is more controversial. It concerns the imperfective aspect, which in English is expressed via the so-called continuous tense.⁵⁰ The imperfective aspect reports a happening as ongoing. But we all know that things may be going on and yet never finish. If Carlos is building a house,

⁴⁹ For reasons discussed briefly in Sect. 2.1.1(iii), we prefer the terminology of wants over intentions at this stage: at least on some understandings of the term, having intentions presupposes having a language to express them, as intentions are often viewed as propositional attitudes. On this understanding, the only literal bearers of intentions are we humans. But we clearly ascribe *wants* to non-linguistic creatures such as cats, and here we wish to target that usage. In the end, this is a merely terminological point. On a sufficiently broad reading of “intention”, one could substitute “intentions” for our “wants” here.

⁵⁰ Languages differ greatly with respect to the way aspect is marked syntactically. In Slavic languages, aspect is indicated directly by the verb form. English has the continuous tense, but German, for example, does not mark the aspect of the verb.

that means that there is a process ongoing whose successful completion would be a real house. But whether that process will be completed successfully cannot be guaranteed. The ground may turn out to be too unstable for building, Carlos may run out of money, or he may change his mind about the house. In all these cases, it will be true to say that Carlos was building a house, but the house was never built. The failure of the inference from “A is φ -ing” to a later “A φ -ed” is known as the “imperfective paradox” (Dowty 1977); it is typical for verbs whose perfective use implies the successful completion of a process that takes some time. While the linguistic data are not conclusive,⁵¹ it may be argued that the imperfective paradox can only arise when the subject of the activity report is an agent. That would provide a further interesting linguistic criterion of agency.

2.3.2 Agency on the Ladder of Attribution

Language, as we just saw, provides some helpful observations about agency, but no fully conclusive criteria. To enrich our overview, the second practice that we focus on is *agency attribution*. As we already pointed out above, it is widely assumed that we have good evolutionary reasons to be competent at discovering other agents and to be rather permissive when it comes to agency attribution. What other agents do is up to them, and they may want things that are not good for us; the fact that they can pursue their aims actively and autonomously makes them hard to predict and potentially very dangerous. And this is just one reason why we attribute agency rather liberally to things around us—such attribution may, for example, also simply be a good heuristic for prediction. Some authors connect this evolutionary argument for a prejudice in favour of agency with a general human susceptibility for superstition, or draw broad conclusions about human beliefs in the supernatural, including religious beliefs (see, e.g., Gray and Wegner 2010; Ma-Kellams 2015). For our purposes, the philosophically interesting issue is not so much how agency attribution evolved or how it is linked to supernatural beliefs, but what we do when we attribute agency. We can learn a lot about the notion of agency by considering our practices of attribution. After our overview of these practices in the remainder of this Section, in the following Sect. 2.4 we will look at cases in which agency attribution goes wrong, i.e., in which we misattribute agency. That discussion will further sharpen the notion of agency for our overall discussion.

⁵¹ Dowty (1977) already mentions a non-agentive example: “The rains are destroying the crops but maybe they will stop before the crops are destroyed” (p. 46). If one takes such examples at face value, one would have to say that the imperfective paradox can occur both for agentive and for non-agentive activity reports—the rain surely isn’t an agent in the sense that we are considering here. But if the rains don’t destroy the crops in the end, is it true that they were destroying the crops? The assessment of competent speakers appears to diverge at this point, while in the case of Carlos’s unfinished house, there is general agreement.

We have already discussed some specifics of agency that are ingrained in the way we talk: natural language incorporates assumptions as to what is special about agency by licensing certain forms of speech for agents, but not for abstract entities or for material or living beings generally. We now focus on another way in which agency shows up as a separate ontological category, namely, in relation to our practices of attribution.

For starters, we can connect to one of our linguistic observations above: Agents *do* things such that what they do is *due to them*, in a sense that goes beyond mere causal attribution. To acknowledge this is important, for example, in social interactions. We may know very well that it was John's foot that stepped on ours in the crowded bus, the causal facts may all be completely obvious—but it will then still be an open question, and important for how we react, whether that causal happening was agentive (a deliberate act) of John or not. Note, also, that even if it was an act of John's and not, as we would say, a mere accident, that does not by itself necessarily settle the further question whether we can attribute this action to John as a *moral* agent, and even if that question has been answered positively, it may still be open whether John deserves to be *blamed* for what he did.⁵²

Our practices of attribution—of saying that some happening was *due to* some entity or other—thus come in different forms that we need to distinguish. As our example shows, agentive attribution is sandwiched between a lower form—mere causal attribution—and a higher, specifically human form—here, moral attribution. This hierarchy is, first and foremost, one embedded in our actual practices. But as with our observations about language, we may hope that our attribution practices, which have developed in interaction with the real world, incorporate distinctions that are usefully related to what there really is.

The Four Rungs of the Ladder of Attribution In what follows we will describe the hierarchy of our attribution practices as a “ladder of attribution”,⁵³ with a nod to Pearl's “ladder of causation” (Pearl & Mackenzie 2018, Ch. 1), which is motivated by the role of causation in advanced AI. Pearl's ladder has three rungs: correlation, intervention, and counterfactual thinking. Of the three types of attribution we have so far discussed—causal, agentive, and specifically human—two fully align with entries on Pearl's ladder, as we will argue: Intervention corresponds to causal attribution, while counterfactual thinking corresponds to agency attribution. In order to fully align the ladders, we need to add a fourth rung in both cases. Pearl's lowest rung, correlation, corresponds to a minimal form of attributing mere coexistence. This lowest rung on the attribution ladder even makes sense for abstract entities.

⁵² His action can even deserve praise—he may have observed a pickpocket targeting us, and his stepping on our foot may have been the best means of attracting our attention in the given circumstances. As so often in life, context matters.

⁵³ These ideas, including substantial parts of the material of Sects. 2.4 and 2.5, were first discussed in the Theoretical Philosophy Colloquium at Zürich (8 December 2017) and in the Ghent-Brussels Seminar in Logic, History and Philosophy of Science (8 February 2018). Thanks to the respective audiences for illuminating discussions.

On the other hand, the specifically human form of agency attribution finds no correspondence in Pearl's original ladder of causation, as it requires not merely counterfactual thinking, but also a reflective form of perspective-taking that is human-specific and which may presuppose having a language. We propose to add an additional fourth rung to Pearl's ladder, which is, we hope, in line with his general approach. Pearl's extended ladder thus has the four rungs of (1) correlation, (2) intervention, (3) counterfactual thinking, and (4) reflective perspective-taking. The corresponding four rungs of our ladder of attribution are, accordingly, (1) the attribution of mere coexistence, (2) causal attribution, (3) agential attribution, and (4) human-specific agency attribution.

To put some meat on the bare bones of our talk of rungs on a ladder, we consider by examples how we can order entities with respect to what can be attributed to them. The ladder reaches from mere abstract assertions of relations of coexistence ("7 is smaller than 8") via physical causal impact ("the cue ball struck the 8 ball") to agency ("Alice cut the cake") and on to human-specific agency, e.g., moral agency ("Bob wronged Alice by breaking his promise to her"). On this ladder of attribution, as we said, agency is sandwiched between mere causation and the human-specific case. By attending to these three respective rungs, we can clarify what it is that we attribute when we attribute agency, and thereby arrive at a clearer notion of agency. The higher and lower rungs provide contrasting examples, while the agency attribution rung provides positive examples, all of which will inform our quest for explication.

In what follows, most of our examples of human-specific agency will be from the moral realm. With a view to our discussion of a phenomenological approach to agency in Chap. 6, however, we also mention a different way of characterising human agency, namely, as *Dasein*. *Dasein* is the term Heidegger (1927) introduces in his phenomenological characterisation of the human way of being in the world. In Sect. 6.2, we will suggest that a basic structure of being in the world can be characterised as acting in the world, corresponding therefore to our rung 3. The specifically human form of acting in the world, *Dasein*, corresponding to rung 4, is then distinguished by a self-referential structure of *care*, as we will discuss in Sect. 6.6.1.

Focus on Ontology Before we get to the details of attribution, we stress again the *ontological* focus of our discussion: We are convinced that agency is not just in the eye of the beholder, as a specific attitude that we can take on select entities in the world—there is more to agency than "the intentional stance" (Dennett 1984).⁵⁴ Relatedly, it will not do, on our view, to take agency to be a socially

⁵⁴ On this issue, too, we agree with the perceptive diagnosis of Steward (2022), who notes that Dennett's view, unlike hers and unlike ours, "is designed to be compatible with the view that, ultimately, everything is mechanism", with any differences to be made constituting "matters of degree, not of kind" (p. 104). See also Steward (2012, Ch. 4.6).—The notion of an "intentional stance" is additionally normally connected to a quite demanding, propositional understanding of intention that we have rejected (again in line with Steward). On such a high-level understanding

constructed notion. Agency is not like money or citizenship.⁵⁵ Certainly, there are many important aspects of agency—especially of human agency—for which social roles and norms play a crucial role. But the notion of agency itself cannot be a social construct, because it underlies the very idea of social construction, which is something that only agents can do. Agency is real, and the concept of agency is a proper subject of ontological, not just of psychological or sociological investigation. At least, such a realist conception of agency, to be further detailed in what follows, informs our investigation. We will therefore not be satisfied with a mere clarification of our attribution practices. We want to understand what is going on in the world, really and specifically, when agency is exemplified. But as a step towards that, it is useful to take a close look at agency attribution. If agency exists, and because agency exists, there have to be proper cases of the attribution of specific happenings, actions, to specific entities, agents—and it will be helpful to get clear on these cases. Furthermore, our investigation will pave the way for a detailed discussion of when our agency attributions get things wrong (see Sect. 2.4). That discussion will actually reinforce our realist stance: to get things wrong in specific cases implies that one could also get things right in others, so that there has to be something, really, that our attribution practices refer to.

The Place of Agency: Rung 3 As we said, the rungs on the ladder of attribution build one upon another, from (1) mere coexistence via (2) causal and (3) agentive on to (4) human-specific attribution, e.g., moral attribution. By discussing these rungs, we can bring out the notion of agency that is at issue here by means of contrast. We can skip rung (1) and start with the second rung. (2) On that second rung, we have causal attribution: certain systems move or change in certain ways that cause certain other things to happen. A stone hits a window and thereby causes the window to shatter, a drop of water falls into the sink and thereby causes a certain noise. These are simple cases of mere rung-2 causation, and there is no agency involved in them. (3) On the third rung, there is the attribution of a more specific happening, an action, to a certain persisting being, an agent. Agents do certain things, as we say, on purpose: things that are up to them and which would not have happened if it were not for them. Instances of rung-3 agency are all around us and clearly within our view. When I brush my teeth, when you walk around, when a bird flies through the city looking for food, or when a cat preys on that bird, all of this is agency. (4) On the highest, fourth rung, we focus on human beings and what they do. Importantly,

of intentionality that presupposes having a language, taking an intentional stance towards a system must in almost all cases be unwarranted. Seeing such high-level intentions at play in the world outside the human sphere must mean to overshoot the mark, or to engage in superstition. To repeat, we aim at an understanding of agency that embraces many kinds of ontologically real animal agency. Cats literally do things, this is not just a useful way of talking.

⁵⁵ These notions have a huge impact on what goes on in our world for sure. But they have that impact only through our social practices that sustain them. There is nothing in those green slips of paper, viewed merely as material objects, that explains why it should be possible to exchange them for a meal.

there is the moral attribution that we normally practice when we interact with our fellow human beings. Strawson (1962) points out that the so-called reactive attitudes—e.g., moral praise or blame, resentment, and gratitude—play a crucial role in our everyday interactions, and that the “objective attitude” of withholding the moral attribution and thereby treating another person as belonging to a lower rung, e.g., as merely a physical cause, is atypical.⁵⁶ Moving down the attribution ladder with respect to another human being is relevant for certain excuses, or for a clinician’s or biologist’s attitude in their studies. In the normal case, however, we do not push other human beings down to lower rungs on the ladder. Instead we generally presuppose the possibility of *morally* assessing what the others do (and also what we ourselves do).

The moral kind of attribution is highly important for us, no doubt, but it is also highly specific. As far as we know, only we human beings are able to use this kind of attribution, and only we human beings are the proper targets of such attribution. (The same holds for *Dasein*.) This need not be the final word, of course. The issue may be broadly empirical, and it is certainly open for discussion. Maybe we will learn things about primates or dolphins or extraterrestrial life forms that will point us in another direction. But at least for now, the safe ground for the highest level of attribution is with us, human beings.

We do use rung-4 moral attribution figuratively, to be sure. We may scold our cat for behaving badly, for clawing our sofa for example, and we may even shout at our evil laptop when it gives us a blue screen. But when you come to think of it, these are not cases of *moral* indignation on our side, really. The attribution of moral agency is in these cases merely figurative, for the sake of emphasis.

2.3.3 A Feline Example: The Agency of Cats

Having delineated human-specific agency (rung 4) from mere agency (rung 3) in the abstract, we now target a specific example, also with a view to adding specific details pertaining to the difference of agency (rung 3) vs. causation (rung 2). Consider, thus, the agency of Tibbles the cat. A cat can do things. He wants certain things and not others. He does certain things in order to achieve his aims. A cat can plan, choose one course of action over another, refrain from following through on an impulse, and more. Unless good reasons to the contrary are provided, we should take all of these ways of talking literally. This is, again, a broadly empirical issue, and there may be reasons to disagree with some of it, but we do not know of any good reasons to disagree with all of this.

⁵⁶ We mention Strawson here because of his clear perception of a relevant difference, but his analysis does not fully align with ours. Strawson in fact throws together our rungs 2 and 3 in his “objective attitude” basket.—The broader context for Strawson’s remarks is his critique of a metaphysical approach to (moral) agency. See Watson (1987) for a critical evaluation of Strawson’s view on that matter.

Cats act, but, as we already said, cats are not proper targets of moral attribution. In the strictly moral sense, cats can do no wrong. Even if a cat hurts another being, e.g., if he kills a bird, we can say and show that we disapprove of this, and perhaps we can say that *the owner of the cat* has a moral responsibility to prevent this from happening, but we cannot blame the cat as a moral agent. On the other hand, it is true that the cat himself deserves our moral attention. It is wrong for human beings to treat cats in certain ways, and the wrongness here is of the specifically moral kind. This singles out cats as *moral patients*, but it does not make them moral agents. Cats need not reciprocate (and of course, being cats, they normally don't).

So much for delineating feline agency "from above": Cats are agents, but they are not *moral* agents. Rung 3, yes, but rung 4, no. To illustrate the difference between rung 2 and rung 3, between causal attribution and agency attribution, let us first assume that Tibbles sheds some hair on your black carpet. This may annoy you (or it may please you—your attitude in this is not important), and if it annoys you, you may try to stop it. (Likewise, if it pleases you, you may try to increase the shedding.) In trying to stop it, you will treat the cat as a physical object with certain causal powers, as an object that causes certain things to happen and not others. If you want to change the cat's relevant causal powers, you can take various routes. You can brush him in order to remove the hair that he would otherwise shed on your carpet, you may perhaps try to change his diet to get him to shed less hair, or you may force him to wear a cat sweater. But you will not, let us hope, try to *train* the cat to shed less hair. That is not something that he is in control of, in the proper agentic sense of the word. You may train him not to go to an area where his shedding hair annoys you, but you cannot train him to alter his physical make-up with respect to shedding hair.

On the other hand, suppose that Tibbles claws your sofa. In this case, it is again possible to view the cat merely as a physical object and alter his physical make-up to make such behavior impossible, e.g., by declawing him. Let us hope that you will not go down this road, because that is at least unfriendly to cats, and on some accounts, even morally wrong. (As we said, the cat is a moral patient deserving our moral attention.) Whatever your own attitude, it is certainly an option to try and *train* the cat not to claw your sofa. This is something cats can learn. There are various ways to discourage them from clawing, or indeed from showing many other kinds of behaviour. Through this teaching attitude towards your cat you are treating him as an agent: as a being that can be trained to do certain things and not others, as a being that can learn. In order to provide a foothold for learning through feedback on behaviour, you will thereby have to assume that the cat wants certain things (e.g., food and attention) and not others (e.g., being sprayed with water), and that he has certain impulses of his own giving rise to situations in which you can train him.

Note that you can take the very same attitude towards other human beings, and also towards yourself. It is not always easy (and it would certainly sometimes be morally wrong) to switch into this mode (Strawson's (1962) already mentioned "objective attitude") because of the morality that infuses all of our human-human interactions. But good and morally unproblematic examples of mere agency attribution towards human beings, of mere training, can also easily be found. Think of

sports. For example, you may be interested in improving your swimming style, and take swimming lessons. You will then get another human being, via normal, morally infused human-human-interaction (e.g., through a contract), to provide feedback on your behaviour, on how to improve certain skills or other aspects of your agency, and thereby to treat you merely as an agent. You may work on your breast stroke, learning to adjust the movement of your arms and legs in swimming, based on the trainer's feedback. As far as we can see, there is no trace of moral attribution in the kind of feedback involved in this training. The trainer will recommend certain types of behaviour and discourage others, in a morally neutral way.⁵⁷

Real life is, of course, complicated. If you are training in a public pool with others around, some of the trainer's feedback may after all be morally relevant—for example, you need to learn not to bump into the other swimmers in order not to hurt them, as that would be morally wrong. And your teacher may also form other broadly moral judgements about you, e.g., that you are a good student, or that you are lazy and that you do not exert the necessary effort to learn. But both these aspects are already on a meta-level with respect to your swimming movements. The physical training takes place at a level on which moral attribution plays no role, and on which the teacher treats you as a mere agent. They can use language to teach you, which makes things a lot easier, and the fact that you are a language user may already be enough to propel you into the realm of beings susceptible to proper moral attribution, as mentioned above—but again this is a meta-level issue that we should leave to the side.

To return to the cat: When you train Tibbles not to destroy your furniture, for whichever reasons, this is a kind of interaction for which you have to treat him as an agent, not as a mere physical being that simply causes certain things to happen. This training exemplifies the notion of agency that we want to target, linked to rung 3 of our ladder of attribution. Agency is a common feature of our world. It is there, all over the place, and it can be clearly distinguished from rung-2 mere causation as well as from rung-4 human-specific attribution.

2.4 Attributions of Agency, Their Defeaters, and Agency Scepticism

Having dealt with our practices of (agency) attribution at some length in Sect. 2.3.2f., we now move to a discussion of ways in which agency attribution can go wrong. Such a discussion is warranted because it delivers on three important

⁵⁷ The fact that the trainer's feedback is morally neutral does not mean that it has to be confined to the mechanics of your movements. Mental imagery plays an important role in training various bodily skills, certainly in adult human beings, and a good trainer will use these resources. The point that is important for our purposes here is that even the use of such sophisticated resources does not move us from agency attribution to moral attribution yet. Thanks to Katia Saporiti for a discussion of this point.

points. First, it rounds off our discussion of agency attribution. By showing that a clear line can be drawn between getting things wrong and getting them right, this affirms our realist stance towards agency attribution. Second, we will be able to link to the topic that came up both via Nagel's worry and in our discussion of the issue of eliminativist reductionism about agency: couldn't it be that the notion of agency itself is incoherent, that the notion comes with demands that are impossible to fulfil, so that the notion cannot be instantiated in our world, on purely conceptual grounds? Through our discussion of when agency attribution fails, we can expose these worries as what we will call *agency scepticism*—as runaway generalisations of initially plausible considerations. We hope that this can help to lay them to rest. Third, our discussion paves the way for the important topic of whether there could be artificial agency. If artificial agency is possible, then this means that the defeaters of agency attribution discussed here do not necessarily apply to would-be agentic artefacts across the board. All together, thus, our discussion will help to further sharpen our explication of agency with a view to providing a target for modelling at the end of this Chapter.

2.4.1 *Specific Defeaters of Agency Attribution*

Suppose you are inclined to attribute agency to a system: you see, or otherwise perceive, something *as an agent*. In some cases, this will turn out to have been a mistake. It was, maybe, just the wind blowing about some dry leaves, a piece of wood floating on the surface of the pond, or the sound of a falling rock loosened by the heat of the sun. What prompts us to withdraw an agency judgement? To put this point more generally, and in useful philosophical jargon: What are *defeaters* of agency attributions? This question provides an additional helpful perspective on the notion of agency, and by highlighting important moves in the dialectics, it also prepares the ground for a detailed discussion of the possibility of artificial agency in Sect. 2.5.

Discussions of defeaters play an important role in epistemology: we can learn a lot about knowledge by finding out in which cases knowledge attributions are defeated and have to be withdrawn. The philosophical literature on so-called Gettier cases provides many studies of knowledge defeaters.⁵⁸ For example, if it is exactly 12 o'clock and you look at a broken clock that shows 12 o'clock, you will form the true belief that it is now 12 o'clock, but arguably, your claim to knowledge of the

⁵⁸ Gettier (1963) describes a classic scenario in which somebody has a justified true belief, but not knowledge, of a certain proposition. In Gettier's example, the knowledge attribution claim is defeated by the deviant nature of the justification of a disjunction. The example given in the main text is structurally somewhat simpler, but it makes the same point. See, e.g., Ichikawa and Steup (2024) for an overview of the discussion surrounding Gettier cases. The terminology of "defeaters" is used widely in current epistemology; see, e.g., the *Synthese* special issue on the topic edited by Moretti and Piazza (2018).

time is defeated by the fact that your method of arriving at your belief (looking at the stopped clock) is not reliable (you would have formed a false belief a minute later, and in fact at any other time).⁵⁹

In moral philosophy, there are also some explicit discussions of defeaters, even though the terminology is not used as much.⁶⁰ Since agency is often not separated from issues of morality, it is perhaps not accidental that there is no explicit discussion of defeaters of agency attributions in the philosophical literature, at least as far as we know. There are indeed many similarities between the agency case to be discussed here and the moral (free will) case, but the moral case is much more complex, and so there are also important differences. We will highlight a few of them in the following, but we do not aim at providing a complete overview of the parallels here.

Defeaters can broadly be classified as practical (such as control, Sect. 2.4.1.1) or theoretical (such as understanding, Sect. 2.4.1.2).

2.4.1.1 Control as a Specific Defeater

The basic idea of control as a defeater of agency attribution is this: What is controlled (or manipulated) is not autonomous, and thus cannot be an agent. Control is a practical obstacle to agency, because the source of what happens lies elsewhere. As there are different kinds of control, it is useful to separate a number of subcases.

Foreign Synchronic Control A strong and uncontroversial defeater of agency attribution is foreign control. If I see a system acting in a way that looks intelligent and goal-directed, I will be inclined to treat the system as an agent, but if I find out that there is another agent (in the cases most easily imagined, a human being) that is really in control of the system and which directs its movements, I will withdraw my attribution of agency to the system, shifting attribution of agency to the real agent behind it instead. The most simple examples would be direct mechanical control, as in a puppet that hangs on strings, or as in the famous chess-playing Mechanical Turk constructed by Wolfgang von Kempelen in 1770. A bit less directly, there is synchronic electrical control by wire, as in operating a crane from the ground. Most controls in cars or in planes used to be direct mechanical, but have meanwhile been changed to direct electrical, or at least to a mixture of both. One step further, there is synchronic remote control, as in toys for children or for adults, such as radio-remote-controlled cars or drones.

⁵⁹ The stopped clock example is Russell's (1948, 98), predating Gettier (1963) by many years.

⁶⁰ One positive example is Part II of Echeñique (2012), "Defeaters of ethical ascription", which concerns Aristotle's discussion of cases in which we withdraw the attribution of ethically significant behaviour to an agent because there was, for example, coercion or force involved. More broadly and terminologically less specifically, the study of excuses, which often defeat moral blame, has been an important topic at least since Austin's classic "A plea for excuses" (Austin 1956).

As distances between controller and controlled increase, remote control becomes less direct—radio waves travel with the speed of light, which is enormous but nonetheless finite. Therefore, long-distance remote control requires a certain division of tasks and thus, some autonomy of the system that is controlled. This is certainly so for automated drones currently deployed, e.g., by the U.S. military. For operations in the Middle East, for example, control centres are located in Europe in order to keep the signal times short, but still a lot has to be going on within the drones themselves.⁶¹ And robots exploring Mars or other space probes, for whom the time delay in communication is even more significant (in the order of minutes or more), need even more on-board autonomy and admit only relatively rare interventions from the outside controller. Still, the type of control that such systems are subject to, at least currently, defeats agency attribution, also in connection with the following issue.

Programming as Diachronic Direct Control When a system is programmed to perform a certain task and then left alone, outside synchronic control in the above sense is not at issue, as nobody intervenes, but still we normally take programming to defeat agency attributions. Many programmed systems lack the flexibility required for agency attribution in the first place, so that we need not even look at their inner workings to reject agency attribution: think of how pathetic a cleaning robot looks if it repeats an unsuccessful attempt at overcoming an obstacle over and over again because it is programmed to do so. But programming need not always lead to completely inflexible behavior; a certain amount of flexibility can be programmed into systems to allow them to break out of unsuccessful routines. All of that depends on the skill of the programmer, who ultimately is in diachronic control of the system since the system only carries out her instructions via its algorithm. Arguably it is this feature of instructions according to a foreign purpose, combined perhaps with the lack of a purpose of its own, that leads us to retract agency attribution from systems that are programmed to do what they do. A challenge to this clear verdict about programmed systems comes from the ever increasing capabilities of systems that are *programmed to learn*.

Foreign Purpose: Diachronic Indirect Control? We are only beginning to harness and to understand the power of machine learning. It is becoming increasingly clear that artificial learning systems can show a large amount of flexibility in their behaviour. Programs such as AlphaZero (Silver et al. 2018) can surprise us through their unexpected solutions to the problems for which they are trained or for which they even train themselves (in the case at hand, e.g., winning at Go). Depending on the learning scheme they exemplify, future artificial learning systems might show really competent behaviour in a wide variety of environments, and perhaps also in interaction with real biological agents, including us humans. But even if systems are built and programmed in such a way that they can learn, there will be a foreign purpose behind their existence and development—their constructor's. That foreign

⁶¹ Ethical questions about who is really in control of what happens immediately spring to mind.

purpose by itself may amount to some kind of indirect control and thus already undermine agency attribution. In these cases, however, the verdict does not seem so clear: A system whose existence is due to a foreign purpose may still be an agent.

Consider, as a parallel, the case of farm animals and pets, whose existence is also due primarily to a foreign purpose—for example, their breeder's. Most of these animals would not be around if it weren't for some specific, mostly economical human purposes. Yet this does not take away the fact that these animals also have a purpose of their own, that there are things that are good or bad for them, that they have needs and, in the case of those animal that are agents, also specific, individual wants (see Sect. 2.1.1.iii). For animal agents (including humans), agency attribution is not revoked if one finds out that they came into existence only because somebody else wanted that to happen. Tibbles the cat is probably around only because somebody wanted a cat, but that does not mean that he is not an agent.

The fact that an entity's existence is due to a foreign purpose *by itself* thus does not amount to a defeater of agency attribution. But if such an entity really is an agent, this implies that there will be two—usually quite different—standards of normative assessment under which the entity's behaviour can be considered. If only one such standard is in play, this may show that agency attribution is defeated after all. To illustrate, it helps to look at behaviour that is out of the ordinary. Farm animals and pets sometimes behave such that they do not fulfil their breeder's or their owner's foreign purpose—for example, some people have dogs for company, but a dog may one day slouch to a corner, curl up and lie down rather than come when called. In this case, we will normally not invoke our standards according to which the dog fails to fulfil her function of being a companion, but invoke the animals' own normative standards when describing what is going on. As a biological agent, the dog can become ill and then, let us hope, recover or be healed. The normative standards set by the purpose of the breeder or the owner, which are responsible for the animal's existence, in such a case are overruled by the animal agent's own normative standards of its health and well-being, and perhaps also of its wants and needs.

If a technological system is really an agent, a similar doubling of normative standards must be in place. After all, such a system's existence will certainly be due to its constructor's purpose—otherwise we would not say that it is a piece of technology. And from that perspective, the system will perform well or badly (function or malfunction), depending on whether it fulfils the constructor's purpose or fails to fulfil it. But as an agent, such a system must also have a purpose of its own, at least in the minimal sense of having certain individual wants (see again Sect. 2.1.1.iii). In case of non-ordinary behaviour, there must be cases in which that agent perspective, rather than the function-malfunction dichotomy, provides the standards of assessment. If there are no such cases at all, we would have to say that agency attribution is indeed defeated—not primarily because of the foreign purpose, though, but because of the failure of a purpose of its own. If only the functioning vs. malfunctioning categories are in play, and the system is always *repaired* or *rebooted* rather than, e.g., being *healed*, this would show that the system is at least treated as

begin answerable only to the standards of its foreign purpose. A necessary condition for agency—having individual wants—would thus be shown to be absent.⁶²

2.4.1.2 Understanding the Inner Workings as a Defeater?

Now that we have discussed the issue of various forms of control and foreign determination of a system's function as possible defeaters of agency attribution, we need to look at a different class of possible defeaters, in which the source of the possible defeat is not, as it were, practical, but theoretical. Are we still warranted to ascribe agency to a system once we *fully understand* its inner workings? Or is such understanding a defeater of agency attributions?

At least psychologically, insight into the inner workings of a system often appears to be a strong defeater. If one can view a system as a mechanism whose inner dynamics one understands, as in bacteria, agency attribution loses its appeal. And in fact, it seems to us that a lot of the current free will scepticism relies on an inference from (purported) knowledge of inner workings of a system to the absence of various kinds of autonomy in that system and, as a consequence, to the absence of free will. At least implicitly, then, it is assumed that knowledge of the inner workings provides a defeater for free will. Systematically, many of these arguments should really be phrased in terms of agency attribution and not in terms of the human-specific attribution of free will and moral agency, since agency attribution is really the level at which the defeaters seem to do their work. If a system does not act, then of course it cannot be a moral agent anyway. It is, therefore, certainly admissible to use arguments defeating agency attribution in a discussion of free will. But it makes good sense to demand that these arguments be discussed in their proper setting, as agency arguments, and not as arguments about free will. In action theory, bringing up moral issues too early generally tends to muddy the waters.

Inner Mechanisms as Defeaters It seems that quite often, it is the mere image of an “inner mechanism” that plays the role of the defeater, independently of any specification of that mechanism in empirical terms, and thus, independently of the degree of our actual knowledge of inner workings. This raises the question of when that image is appropriate, and when it should be rejected.

The case of bacteria, briefly mentioned above, seems to be one for which the image of an inner mechanism as an agency defeater is perfectly apt. Look at the behavior of bacteria swimming around in a liquid in a way that appears to be purposeful and goal-directed. Their movement is regulated by chemotaxis, which brings them close to an attractant by sampling the attractant concentration and regulating their movement accordingly, resulting in a form of a biased random walk. Chemotaxis is part of the bacteria's life processes. It seems apt, however, to describe bacterial chemotaxis as a *mechanism*, which is furthermore quite well understood

⁶² We do not yet have much data on how we human beings will behave in interaction with such future systems, and what our attitudes will be. For one relevant study, see Straub (2016).

even at the molecular level by now.⁶³ The self-movement of bacteria, we should say, is therefore mechanical and not agentive. That mechanism, like other aspects of the bacteria's life processes, is based on the use of indeterminism. It is therefore correct to say that a bacterium has various possibilities for its movement—but it is merely figurative speech to say that the bacterium “decides where to swim”. The bacterium *is* its body, it does not *have* a body that it can move agentively. There is no decision and no agency in bacteria, and we know this because we understand their inner mechanism.

Inner Mechanisms in Higher Animals? We have no fully detailed grasp of the self-movement of higher animals.⁶⁴ Still, we do have a good grasp of some of the inner structures that play an important role in their behaviour. We understand the mechanical set-up of muscles down to the molecular level, the coupling of muscles to nerves, the mechanism of signal transduction in nerve fibres, most of the workings of synapses between neurons, and thereby also some of the basic building blocks of brains. By focusing on such a level and labelling it as mechanical, we can cook up arguments that may appear to defeat agency attribution. For example, for any attribution of agency to a person (as in “Alice is raising her arm”), we can provide the outlines of a story about the mechanism of signal transmission in the neuronal system that finally triggers the arm's movement. And once we point to that mechanism, it may seem that the agent disappears.⁶⁵ But this move by itself is not conclusive—it begs the question whether a full mechanical explanation is really available.

There seems to be a certain tension: a true, low-level mechanical explanation of a system's behaviour may indeed work as a defeater of agency attribution, but such explanations are rare, and most mechanisms offered in explanations pertaining to natural systems are rather coarse-grained, incomplete sketches. This is different for artificial systems that we have designed and whose make-up we therefore understand completely—we will discuss the special case of understanding an artificial system as a possible defeater of agency attribution in Sect. 2.5. It is

⁶³ The detailed analysis of chemotaxis shows that it is not as simple a mechanism as one might think. Even bacteria have some form of memory and can make use of environmental noise; see Vladimirov and Sourjik (2009) for an overview. Our main point is, however, not threatened by these additional complexities.

⁶⁴ This might be a merely empirical limitation, which we might be able to overcome in a few decades. There might, however, also be fundamental obstacles barring the kind of full grasp that it seems we have in the case of bacteria, since higher animals are much larger open systems. We will not speculate about this here.

⁶⁵ This, then, is the “zooming in” worry of the disappearing agent mentioned in connection with Nagel (Chap. 1, footnote 1). In the philosophical literature, the worry comes in two variants, depending on whether the mechanism is described in physical or in psychological terms. Even if the agent is described not at the level of a physical or neural mechanism, but at the level of a mechanism involving psychological events (intentions, beliefs, desires), it may still seem that the agent disappears from the scene. This worry is actually the main theme of the paper by Velleman (1992) already mentioned above.

interesting, however, that a reference to an underlying mechanisms, even if that reference is rather sketchy, appears to work as a psychological defeater of agency attributions quite generally. It seems likely that the reason for this lies in a certain general sceptical tendency that is not tied to a detailed understanding of a given system at all, but maybe to a deterministic image of mechanisms operating like clockwork. That sceptical tendency may be connected with the general attitude of asking the agency question with a sceptical sneer (see Chap. 1, footnote 2), but such a connection is not necessary. In criticising the sceptical tendency, we definitely do not mean to be accusing those who arrive at it of insincerity.

2.4.2 *General Defeaters: Sceptical Doubts*

Before the background of the above discussion, Nagel's worry and the related worry of the disappearance of the agent can be seen as steps on the way to agency scepticism. The sceptical argument is that some defeaters of agency attribution must apply completely generally, independent of context, so that agency attribution always turns out to be wrong. This would expose the notion of agency as useless, and perhaps that uselessness could even be linked to the notion's ultimate incoherence.

Do Defeaters Generalise? Which defeaters can be generalised to apply across the board? It is clear that the practical defeaters discussed in Sect. 2.4.1.1 cannot become completely general, as they presuppose manipulation, control, programming, or foreign purpose due to *another agent*. Without that other agent, the practical worries cannot arise, and so, they remain local and always tied to a concrete context.⁶⁶ Theoretical defeaters, such as a detailed mechanical understanding of the workings of a system, however, do admit of a generalised form. Due to our cognitive and other limitations, a full understanding of the inner workings of all agents is of course out of the question, but one might hold that there surely *is* some such inner mechanism at work in all the agents, known or unknown, and that already this fact, or the related "in principle" availability of an inner mechanism account, defeats agency attribution. The idea is that what is really going on, irrespective of whether it is known to us or not, is just natural happenings, ultimately at the physical level, and agency disappears. In a phrase mirroring Nagel's worry, we can say that everything that I do is constituted by things I don't do, because I am made up of smaller parts that interact mechanically. If this line of thinking was correct, it would show that agency is a useful concept for us only because we are ignorant of the basic facts and

⁶⁶ Leaving aside considerations of theological determinism (see note 68), the only way to generalise them would seem to be to let "mother nature" do the controlling by enforcing the laws of nature as a kind of program. This, however, basically amounts to what we describe as a generalised *theoretical* sceptical strategy below. For example, Pereboom's (2014, Ch. 4) "manipulation argument" regarding determinism, despite being directed against free will rather than agency, falls in that category.

incapable of understanding nature at a deep enough level. Accordingly, the notion of agency could really only have an epistemic, not an ontological status.

In our discussion of ontological levels in Sect. 2.2.3 we already pointed out, however, that this line of argument is too quick: the mere fact that higher-order processes are realised via lower-level materials does not prove that there is a lower-level mechanistic account of what goes on at the higher level. Such an account is of course not absolutely excluded either; it may well be so in some cases. Importantly, it *is* so in Conway's "Game of life", a deterministic cellular automaton that can show interesting, apparently higher-order features and which is often used by reductionists to support their claims.⁶⁷ But it is exactly the *determinism* of that system that makes it inadequate as an analogy for agency, let alone free will. In reality, the lower levels of organisation in animals are indeterministic and leave causal slack for higher-order organisation, and therefore, a lower-level mechanistic account of agency is not plausible. Understanding of inner workings may defeat some attributions of agency, as we argued in the case of bacteria, but this does not generalise across the board.

Completely General Defeaters There is, however, yet another approach to positing general defeaters of agency attribution. Curiously, this strategy can only be framed as the existence of *completely general* defeaters. It does not admit of local distinctions like the practical and theoretical defeaters discussed above. Nagel's original worry points in exactly such a direction. Note that the very form Nagel uses is already that of admitting a defeater: The worry is that "[e]verything I do"—an agency ascription—"is part of something I don't do, because I am part of the world"—whereby the agency attribution is revoked (Nagel 1986, 114).

The worry of losing agency by "zooming out" and considering the contributions of the would-be agent as a part of the overall course of natural happenings is suggestive. By itself, this does not yet provide an argument against agency. But the worry can be posed as a dilemma. It comes in two more specific variants, depending on which side one takes in the metaphysical debate about the open future, about the issue of determinism vs. indeterminism. In discussing Steward's list of features of agency in Sect. 2.1.1 above, we already voiced agreement with her agency-incompatibilism: There can be no agency under determinism. After all, if one assumes determinism, the development of the whole universe for all future times is fixed—determinism is the thesis that there is always only a single possibility for the future to turn out, and this must hold no matter how far back in time one goes. Now consider a would-be agent, α . Anything that happens in α 's life time was fixed already 1000 years before α was born (or made): at that time, there was, after all, only a single possibility for the whole future to turn out, which includes α 's lifetime (or period of operation). But α cannot make a difference to what happens in its life time, because that is already uniquely fixed, and so α never settles any open matters, and so α cannot be an agent after all. This is the basic form of Van

⁶⁷ See, e.g., Dennett (1991, 2003).

Inwagen's so-called consequence argument (Van Inwagen 1983).⁶⁸ There is a huge and controversial literature on the consequence argument and related deterministic global defeaters of agency (mostly discussed in terms of free will). Give or take a few details, we hold that the consequence argument is valid: In a deterministic world, there is just no wiggle room for agents, natural or artificial, to make a difference to what happens. And as we argued in Sect. 2.2.3, the whole idea of real ontological levels in nature needs lower-level indeterminism. In a deterministic world, there can be no ontologically real higher-level structures at all, and therefore, no agency in particular.

The other horn of the dilemma assumes indeterminism, i.e., the thesis that there are multiple possibilities for the future to turn out. As we said, indeterminism is basically good news for ontological pluralism and thus, also for agency. There are, however, various arguments purporting to show that under indeterminism, anything an agent appears to do comes down to base-level chance, and by definition, nobody is in charge of base-level chance. So the would-be agent could, again, not *do* anything, and thus, could not be an agent after all; it could only be tossed around by chance happenings that make no sense to it. Discussions of this worry have also created a huge literature. For our take on the issue, see Müller and Briegel (2018), which constitutes an earlier, compact attempt to make sense of agency under indeterminism. In that paper, we already use agency modelling via Projective Simulation, which in this book will be discussed in detail in Chap. 5.

It seems indeed that the deterministic horn of the dilemma is inimical to the existence of agents. We gave some reasons for that above. But given that there are agents—for example, we are agents—it cannot be that both horns of the dilemma rule out agency. So, from the existence of agents and the incompatibility of agency with determinism, we have another argument for indeterminism.⁶⁹ And in our view, indeterminism indeed leaves room for agency. It is, to be clear, not enough by itself: Just given indeterminism, nothing about agency can be deduced. Agency needs structures in nature that build upon nature-given basic indeterminism such that new ontological levels are really established. Such higher levels did, for all we know, not exist in the very early universe, when the relevant material wasn't yet available, as it was simply too hot. As we pointed out in Sect. 2.2.3.4, it took a while for even basic

⁶⁸ Van Inwagen, like most others, adds a reference to states of the universe and to laws of nature in his account of determinism. That additional layer is omitted here because it is unnecessary and indeed beside the point. Determinism is a metaphysical thesis about the absence of possibilities, and not about inferential relations between descriptions of the universe's state together with expressions of laws of nature (Müller & Placek 2018; Steward 2021). Be that as it may, the general form of the consequence argument can be found in many places and in different forms. Historically, some important arguments along these lines were based on theological discussions about divine omniscience vis-à-vis an open future leaving room for human beings as true moral agents. See, e.g., Øhrstrøm and Hasle (2024) and Øhrstrøm and Jakobsen (2022) for historical details.

⁶⁹ Many philosophers seem to worry that this reasoning is somehow too quick. Steward (2020) provides an illuminating diagnosis of the dialectics involved here. We agree with her that the argument is sound.

nuclear physics to get going. And our own agency, as well as the agency of cats, arose on this planet only based on additional levels reaching up to life and organised life, which took a long time to evolve. Nothing of this is automatic or in any way guaranteed. But the fact that it did happen proves that it was naturally possible. If it weren't, you couldn't be reading this book now.

Much more would, surely, need to be said. And it is a fact that there are full-blown agency sceptics out there, who hold that the mentioned dilemma (or some more sophisticated variant of such a dilemma) indeed rules out agency. (As we mentioned, given the scarcity of discussions of agency as a separate topic, most of the literature targets free will, which is a higher-level phenomenon concerning *moral* agency. But most arguments really address the ontological level of agency, not moral agency.) Whatever the merits of these arguments, we insist that they be discussed as what they are: general sceptical arguments. They have to apply globally, they cannot be used to adjudicate between good and bad, successful and unsuccessful uses of the notion of agency. This puts them in a different category than the “local” defeaters discussed in Sect. 2.4.1 above.

One prominent defender of a global sceptical argument is Pereboom (2001, 2014), who explicitly calls his position “free will skepticism” (Pereboom 2014, 4). In our view, his arguments against the possibility of agency under indeterminism are not successful.⁷⁰ Successful or not, however, sceptical arguments are important in philosophy, and they are not without their merits. Not every occurrence of a sceptical argument means that the person in question has dropped their “straight face” and put on a “sceptical sneer”.⁷¹ Sceptical arguments can show that there is something problematic about important basic concepts, and they force us to sharpen our ideas. The notion of knowledge in epistemology is a case in point: discussions of sceptical scenarios and arguments for different forms of scepticism have enormously enriched our understanding of that notion and of our ways of using it.⁷² But it does not follow that many people are really convinced (let alone reflect in their behaviour) that they do not know anything after they have read some heavy epistemological discussion of scepticism.

With respect to agency and free will, however, the force of sceptical arguments seems less limited. Indeed there appears to be something like a consensus among the scientifically minded that free will cannot exist and has to be an illusion.⁷³ Arguments that operate on the global level of determinism or indeterminism, which feature prominently in that area, should give the same verdict for agency that they give for free will. By parity of reasoning, thus, the scientifically minded should also believe that agency cannot exist. This is, however, really untenable. Clearly there

⁷⁰ See Müller (2023) for one relevant specific objection that targets Pereboom's so-called “wild coincidences” argument.

⁷¹ See footnote 2 in Chap. 1 above.

⁷² See, e.g., Lewis (1996) for a discussion of a general recipe for cooking up sceptical scenarios and how to react to that.

⁷³ See, e.g., Wegner (2002).

are agents in our world—the scientifically minded at least have to admit that there are agents developing the models and running the experiments on which science is based (see Sect. 4.4). The interesting question, it seems to us, is not whether there are agents, but which systems are indeed agents, and which are not, what distinguishes them, and how those that *are* agents do the trick.

2.4.3 Conclusion on Defeaters

There are different ways that agency attributions can be defeated. Some of the defeaters start from practical assumptions about control, others start from theoretical assumptions about a system's inner workings, and yet others start from global assumptions about the world as a whole. It seems that the latter type of sceptical arguments, while probably making up the bulk of the philosophical literature on defeaters for free will (or for agency), are the least relevant for understanding agency. In particular, global sceptical arguments do not teach us anything specific about the real possibility of natural or artificial agents. Defeaters of the other two types appear, however, to be especially relevant for the latter question. In the following Sect. 2.5 we will go on to discuss whether practical or theoretical defeaters apply to artificial agency generally, without a specific reference to details of the case. This would contrast them with natural agency, for which we know from the examples of human and feline agency that defeaters cannot apply generally.

2.5 The Possibility of Artificial Agents

The main question we wish to tackle now is this: Could there be artificial agents? Opinions are certainly divided. Steward (2012, 91) makes the following useful general claim about artificial agency:

[...] there is nothing in the *concept* of agency itself, I believe, that rules out the possibility of its instantiation in artificially created, non-biological matter.

Steward thus affirms the *conceptual* possibility of artificial agency (even though she is more sceptical about its *empirical* possibility). In her quote, she points to two aspects of artificial agents: First, they will be (as the name says) artificially created entities, i.e., artefacts. Second, they will consist of non-biological matter. The latter point is perhaps a bit tricky—certainly it is true that current efforts in AI that may target artificial agency concern implementations *in silico*, but technologies for manipulating biological matter for various purposes are also becoming more and more powerful, and maybe it is too early to say which types of matter will be the

basis for the first artificial agents (if ever there will be any).⁷⁴ But the point Steward is making seems valid: there appears to be nothing about the specific causal powers (dispositions for physical and chemical interactions) of biological matter (proteins etc.) that establishes that any possible agent needs to consist of exactly that kind of matter. Other kinds of matter have different causal powers, to be sure, but nothing we know about physics or chemistry or agency suggests that complex behaviour such as that of agents cannot be realised in non-biological matter. This seems to be a fair assessment of our knowledge of the basic properties of matter. It is true that carbon-based so-called organic chemistry provides for a stunning variety of possible compounds, and more specifically, that amino acids are really versatile basic building materials. But in life on earth, chemical elements other than carbon and substances other than amino acids also play important roles, and quite generally, we have learned to engineer many different materials, becoming ever more skilful at handling matter down to the atomic scale. So indeed, nothing about the intrinsic properties of biological matter appears to support a claim to exclusiveness as a basis for agency.⁷⁵

This part of Steward's claim is, at any rate, mostly empirical: for all we know, matter that is different from what constitutes life on earth may still provide a basis (a hardware platform, as we might say) for agency. The conceptual claim she expresses thus gets its bite from the fact that artificially created agents must be *artefacts*. The claim is that it is conceptually possible for artefacts to act, and this claim is non-trivial. In what follows we will confront the claim with the possible defeaters of agency attribution discussed in Sect. 2.4. Our discussion will in the end support Steward's claim of the conceptual possibility of artificial agency.

2.5.1 *Generic Defeaters for Artificial Agency?*

There certainly are actual attributions of agency to artefacts that are actually defeated. In fact, given the current state of technological development, it seems fair to say that currently no such attribution is really successful. But this might be just a contingent fact having to do with the level of technological development that we have reached by 2025. The fundamental worry (or, depending on your attitude, the hope) regarding artificial agency runs deeper: there might be defeaters afflicting any would-be attribution of agency to an artificial system. The question we need to

⁷⁴ An additional fine point is whether completely artificially created systems made of proteins etc. would still count as being made of "biological matter". This depends on whether one reads the "bio-" in "biological" strongly, as presupposing life, and then, on whether artificial life is a conceptual possibility. (As already mentioned, Thompson 2008, Part I gives reasons for a negative verdict.) For our discussion, we will not need to resolve these issues, and we can read Steward as referring just to a certain type of chemistry.

⁷⁵ Furthermore, as Van Lier (2025) argues, artificial agents might also be embodied in radically new, e.g., spatially distributed ways.

answer is, therefore, whether any attribution of agency to an artificial system (such as a robot) could possibly be upheld in view of the various defeaters discussed in Sect. 2.4 above. We will not consider the general sceptical worries of Sect. 2.4.2, as they would affect agency across the board. As we pointed out, given the reality of agency, we reject agency scepticism. Instead, we consider in turn those defeaters that we called practical, which are based on control (see Sect. 2.4.1.1), and those that we called theoretical, which are based on understanding (see Sect. 2.4.1.2). Both may be pertinent and may affect artificial agency in a specific way.

Practical Worries On the practical side, the worry is that artificial systems, being built by us for specific purposes and with the aim of fulfilling specific functions,⁷⁶ will lack autonomy. Already with respect to the current state of technological development, it seems that some of these worries can be laid to rest. First, artificial systems need not be controlled in any strict sense. They need no strings, wires, or radio links to steer them by, and they may be able to act fairly autonomously for a long time. The supply of energy will probably require a specific technological infrastructure for the foreseeable future. But then, we, too, rely heavily on a fairly sophisticated technological energy infrastructure without which our lives becomes difficult and dangerous. This can be seen in any power failure, and it also becomes clear when we reflect on the way we come by the food we eat.⁷⁷ Second, strict programming of a system's behaviour is not required either, given that it is possible to engineer quite powerful learning systems whose behavioural routines are not pre-programmed. Another relevant worry from the control side is that the purpose or function of the artefact in question will reside with us. But as discussed at the end of Sect. 2.4.1.1, this is so for many animals as well, where it does not prompt us to withhold agency attribution to them. So it seems that the practical worries against the existence of artificial agents mentioned so far, which target their autonomy, can be laid to rest.

The remaining practical worry is that in contradistinction to the case of animals, in artificial systems there is simply *no* purpose or function that belongs to them *independently of us*, so that their would-be actions cannot *make sense* for them. This criterion is a bit vague, but the total absence of sense-making would indeed be a defeater. While this worry is serious, here we will leave it to the side, because it seems too difficult to make it precise. One version of the worry that *is* precise should, we believe, be dismissed. If one assumes that it is conceptually necessary that an artificial system cannot act in a way that makes sense for itself, then indeed, there can be no artificial agents whose actions make sense for themselves. But why should we believe the premise? It seems obvious that we have to admit that “making sense”, intrinsically purposeful behaviour, has a foothold already at a relatively low level in the animal kingdom, and that has come about through natural developments

⁷⁶ These functions need not be overly technical. Many artificial systems are built just to entertain, to play with, or for aesthetic pleasure. So “function” is to be taken quite widely here.

⁷⁷ Very exceptional circumstances aside, we human beings need a lot of cooperation to sustain our lives. Absolute autonomy is mostly a fairy tale, or rather a Robinson-style adventure story.

building upon natural resources. It is an emergent property of a fair number of natural systems that their behaviour can make sense to them and can contribute to what is intrinsically good for them. What can emerge from what depends on causal powers, and above we agreed with Steward that there is nothing in the causal powers of biological matter that supports an exclusive claim to basing agency. So, at least tentatively, the door seems to be open for sense-making to emerge in non-biological, artificial systems as well.

The sense-making required for *agency* furthermore need not include everything that is implicated in the case of life. Artificial agents, as we tentatively said, need not be alive. Thus, reproduction and the sense-making that comes through caring for one's offspring may be absent.⁷⁸ It seems plausible to assume that the natural emergence of sense-making that we see in biological entities *did* depend on reproduction and selection. This would, however, still leave it open whether sense-making in artefacts must also come by that route. There may be other, more individualistic routes as well. At any rate it seems that the most forceful threat to the possibility of artificial agency is not practical, but theoretical, and so this is what we will focus on.

Theoretical Worries On the theoretical side, the main worry is this: For systems that we engineer, we will have a full understanding of their inner workings, since these inner workings were made by us according to our plans. Anything that we make, that is, seems to embody a mechanism that we designed and which we fully understand. And once we are able to see a system's behaviour as mechanical, we have good reasons to withdraw agency attribution, or at least to view such attribution as merely a pragmatic reaction in the face of high complexity, but not as indicative of something real and ontologically special. Knowledge of a base-level mechanism, as we said in Sect. 2.4.1.2, is a powerful defeater of agency attribution.

The task before us, as we see it, is to explore this worry with regards to a realistic picture of future autonomous artificial systems that takes into account their development over time and, most notably, their nature as learning systems. The question is whether agency can emerge in a system for which we have a model whose basic constituents we fully understand, but whose development over time may be autonomous and open, as it can make use of naturally given indeterminism for its organisation.

It seems helpful to discuss this worry in an abstract form that focuses on full understanding, rather than on the fact that we made the system. The worries arising specifically from our having made the system are, we believe, mainly practical, having to do with the idea that we control what we make, or that what we make cannot have a purpose to itself. As we said above, these worries do not seem decisive. But we need to tackle the theoretical worry in more detail.

⁷⁸ See the opening discussion of Sect. 2.3, esp. page 44.

2.5.2 *Artificial Agency: Attribution, Ignorance, and Emergence*

Quite generally and abstractly, our issue is this: Given that we understand what a system is made of and how its parts work, how can there be interesting emergent behaviour in the system as a whole? And can such emergent behaviour amount to agency? We will ignore the source of our understanding of the system for now—it might be due to our detailed investigation of what we found in nature, as in the case of bacteria, or due to the fact that we brought the system into being by constructing it from its parts according to our plans in the first place, as in the case of artefacts.

The case of bacteria is helpful to illustrate a challenge having to do with ignorance. We can understand the behaviour of bacteria (at least their self-motion) via the mechanism of chemotaxis, which defeats the attribution of agency to them, as we said in Sect. 2.4.1.2. We understand too well what is going on in bacteria for the concept of agency to gain a foothold. It is important to note that this understanding does *not* operate at a putative lowest level of fundamental physics, not even “in principle” (even though some philosophers or scientists of a reductionist bent will be quick to claim that as well). The matter that is in this case mechanically used as material is in part biochemical, in part it consists of already organised biological subsystems (like the flagellum that propels the bacterium forward by rotating clockwise or counterclockwise). It is our actual understanding *at that concrete level* that provides an actual agency defeater. Ontologically, bacteria belong to the level of living beings, but not to that of living agents.

The worry that comes up at this point is the following. The contingent availability of our current understanding (our having advanced the science of bacteria far enough) should not make a real difference as to whether a bacterium or some other single-celled organism is an agent or not. At a time before we gained the relevant understanding, we may well have attributed agency to them. For a case in point, consider seventeenth century descriptions of animals discovered under the microscope, e.g., by van Leeuwenhoek.⁷⁹ These agency attributions have been shown to be mistaken by our later understanding, and they have been revoked: As we said, we agree with Steward (2012) that single-celled organisms are not agents. Ignorance may provide a favourable *epistemic* environment for agency attribution, but it is not a stable ground on which such an *ontological* attribution can rest. Now it is a fact that animism and anthropomorphism tend to be gradually diminished as scientific understanding is growing: more and more attributions of agency have had to be revoked. Apart from understanding the movement of bacteria, we have also pretty much gotten rid of ghosts. So don’t we have to extrapolate, and assume that “at the end of science” we will have a full understanding of all the mechanisms in

⁷⁹ See Wellmann (2024, Ch. 5) for a detailed account. For example, van Leeuwenhoek describes the human spermatozoa that he discovered as homunculi, little human agents.

nature, including those that give rise to what we take to be our own agency—which would show that *all* agency attributions are ultimately defeated?

This thought is based on a slippery slope of moving from specific defeaters in relatively simple cases to a global, sceptical assumption of general defeat. We know that this cannot be right—there are agents, and thus there are agency attributions that are not just currently undefeated, but really undefeatable. To repeat, cats are agents, and we are, too.

There is, however, a remaining challenge, as the reflection on the dynamics of ignorance seems to leave only two ways to go forward. One is to hold on to a strong link between agency attribution and ignorance: full understanding of inner workings truly blocks agency attribution. As realists about agency, we have to defend undefeatable agency attributions, but if full understanding blocks attribution, such realism would come at a substantial price: it would be necessary either to posit fundamental limits to our understanding of the inner workings of natural systems, or to claim that addressing a system as an agent precludes certain scientific investigations of that system, and vice versa. Both ideas are unappealing. Note also that they would probably establish the conceptual impossibility of artificial agents: Since we understand their inner workings by assumption (after all, we made them), such systems could not be agents, by conceptual necessity.

The other way forward appears to be much more promising, and it connects well with our discussion of ontologically emergent individuals above (Sect. 2.2.3). As we argued there, even in a system whose parts we fully understand, there can be novelty and emergence. To take up our main example: the sun is an entity whose inner workings we understand very well, but still the sun is not merely matter arranged in some special way. The sun is real. In a similar way, a bacterium is a real living entity; the fact that it is alive is not diminished by our understanding of its inner workings. And if we understood a cat at the respective level of physiological detail, this would not take away her life, either. For the cat, it would also not take away her agency, because such understanding would—in contrast to the case of the bacterium—not amount to a complete mechanistic understanding of the cat's bodily behaviour and its self-motion. Even if we understand the processes by which the cat is alive at a level that exposes many of them as quasi-mechanical, the cat's organisation as a living being leaves open manifold possibilities regarding the motion of her body that the cat *as an agent* can organise, settling what to do in a specific situation. Even given a fully detailed physiological understanding of what goes on in the cat, she could (and would) still surprise us.

This, then, is our answer to the theoretical worry regarding artificial agency. The real existence of ontologically emergent individuals is not threatened by an understanding of their material basis, as long as that basis leaves open a relevantly large enough range of possibilities for the emergent individual to organise. Thus, a full understanding of an agent's body is no threat to its agency, given that the agent's body still offers open possibilities for the agent to settle. Agency attribution need not be backed by ignorance of inner workings.

2.6 Our Explication of Agency

Based on our discussion of ontological levels, agency attribution, and other aspects of agency, we are now in a position to provide an updated explication that expands on our discussion of Steward's definition of agency in Sect. 2.1.1. As we said there already, we differ from her approach by emphasising the diachronic aspects of agency and of an agent's individual development during its lifetime (or period of operation).⁸⁰ Like Steward, we structure our explication into a list of items that we take to be necessary for agency, but with no claim to these items' independence. Indeed, we will point out some interdependencies along the way.

2.6.1 *A List of Four Criteria for Agency*

Very simply put, an agent is (i) an embodied, concrete material individual that (ii) can settle matters regarding its body, which (iii) wants certain things for itself, and (iv) whose history is one of learning, developing an individual memory, and adapting to its environment. In virtue of these conditions, agents must be ontologically emergent individuals according to the three criteria detailed in Sect. 2.2.3.4: (1) Their matter—their bodies—must leave open possibilities for the agents to settle (see items i, ii); (2) agents must organise their bodies with a view to what they want (see item iii), and (3) they have to fit into their environments (see item iv). For the latter task, agents are specifically equipped because of their capacity for learning and adaptation.

i. Embodiment An agent is a concrete material entity with certain causal capabilities for making a difference with respect to what goes on in the world. Already by this minimal notion of embodiment as concrete material existence, an agent comes with a basic notion of historicity—concrete material entities begin to exist at a certain point in time, through various processes that are specific for the kind of entity in question, then they persist over time for shorter or longer periods, undergoing a variety of (again, kind-specific) possible changes, and then their history can end in several ways that are also different for entities of different kinds. Agency, however, comes with a more complex notion of embodiment. First, like the sun and like living beings, agents are open physical systems whose persistence is marked by the maintenance of a certain kind of dynamic equilibrium, and their existence is only possible in environments that are favourable for their specific type of persistence—for example, their material composition and the material composition of their environment has to provide enough open possibilities for maintaining the dynamic equilibrium in question. And while (for all we know) agents do not need

⁸⁰ What we lay out here is in part based on the explication of agency we provide in a previous paper (Müller & Briegel 2018). See also Briegel and Müller (2015).

to be living beings (see Sect. 2.5), like living beings, they need to be embodied in a way that organises the matter they are made from. Thus, agents are ontologically emergent individuals. Second, their bodily organisation in turn has to leave open enough possibilities for the embodied agent to make a difference (see (ii)). Third, as complex systems that are set off from their environment, it makes sense to associate agents with an input-output structure in the interaction with their environments, i.e., to identify in them some sensors and some actuators. Through some of these actuators, agents must be able to move (parts of) their bodies.

ii. Settling, Indeterminism, and Agency-Incompatibilism In Sect. 2.2.3 we argued that the existence of any ontologically emergent individual presupposes indeterminism: the higher-order (emergent) organisation has to be based on lower-level open possibilities for the organisation to have any real effect, to be more than a manner of speaking. In the case of agents, we agree with the general suggestion of Mulder (2021) that a higher-order organisation such as agency must be a way of organising something—the agent’s body—that is itself *already* organised. This accords well with Steward’s claim that *agents settle matters concerning their bodies* (see Sect. 2.1.1.iv). These bodies themselves must then already be organised as emergent individual entities, meaning that their very existence already presupposes lower-level indeterminism. But the ontological emergence of the agent is based not on that lower-level indeterminism directly, but on possibilities that are left open for the organised body itself. As we said with reference to bacterial chemotaxis, for some living beings—such as bacteria—the build-up of ontological levels stops there. Bacteria successfully organise lower-level indeterminism, whereby their bodies are real emergent entities, but these bodies appear not to leave open any meaningful possibilities that could be *further* organised, and this prevents them from being agents.⁸¹ Thus, we fully agree with Steward’s picture of agency as settling matters concerning the body, but we wish to add that this actually presupposes *two* levels of indeterminism that build one upon the other: an agent’s settling matters concerning its body presupposes that body leaves open meaningful possibilities to be settled (the agent *has* a body, it is not *identical to* its body), but the very existence of an organised body *already* presupposes indeterminism of the underlying matter that the body organises as *its* material. That is, the incompatibility of agency and determinism is actually at least two-fold, at the level of the underlying physical matter *and* at the level of the body.

iii. Wanting. We said that it is specific for agents that they *want* things. We commented on our choice of terminology above (Sect. 2.1.1.iii)—in the end it does not matter if we speak of wanting or of intentionality, if the latter notion is taken lightly, but quite often it isn’t, so our terminological choice seems safer.

⁸¹ The mere addition of a randomisation device will not change that verdict, as it only provides open possibilities without a way for their higher-level organisation. An organised body enslaved by chance is not in a better position than an organised body subject to a deterministic mechanism. This parallel points to the truth behind the otherwise faulty dilemma discussed in Sect. 2.4.2.

Some things that biological agents want are pretty much uniform for agents of a specific kind. One reason for this is that it makes good evolutionary sense to assume that agents have adapted such as to want what they need, because otherwise they would not be around any more, and what an agent needs depends on the kind of entity it is.⁸² Thus, for example, hedgehogs need a place to shelter to sleep through the winter, and so in fall one sees them *trying to find* such shelter, whereby one sees them *wanting* such shelter. (As Anscombe (1957, §36, italics in original) usefully remarks, “[t]he primitive sign of wanting is *trying to get*”, and we can surely see *that*.) But other things that individual agents want (which, by Anscombe’s remark, we can also often perceive directly) are not uniform among agents of the same type, and very often these are connected with their individual history. I want coffee for breakfast, as I am used to that; you may want tea.

Now, whether kind-specific or individual, what an agent wants has to make a difference for it; there must be something “in it” for the agent. If the agent gets what it is trying to get, that must make a difference for the agent. Wants cannot be purely extrinsic, they have to be tightly integrated with what the agent *is*. It is not easy to provide a criterion for when that would be the case. But to give one relevant indication: in most current reinforcement learning algorithms, the reward function is defined in an extrinsic way, setting a task for the learning agent but not literally constituting something that the learning agent wants *for itself*. One obstacle to building artificial agents is surely the necessary deep integration of things they *really want* into their overall organisation.⁸³

iv. Historicity, Learning, and Individuality as Diachronic Aspects Of Agency

We said under (i) that, being ontologically emergent individuals, agents already come with a notion of historicity at least in the sense that their existence is temporally constrained. As we argued, however, the historicity of agents must be significantly more structured: for example, agents learn, they adapt to their environment, and thereby they acquire an individual memory and individual traits (including individual wants).

The historicity and individuality of agents is connected with two important aspects of agency: agents as *settlers* (see (ii)) and, building on that, agents as *learners*. The behaviour of agents has to be non-rigid, as agents can settle matters left open by the organisation of their bodies in different ways. We already quoted

⁸² In us humans, as we all know, the association between needs and wants has become much weaker, and most—if not all—of us want many things that we do not need (including things that are really not good for us). That complication is one more indication of the complexity of human agency. We will discuss a few relevant aspects from a phenomenological point of view in Sect. 6.6.1.

⁸³ Some reinforcement learning frameworks have introduced formal notions that are somewhat related to wanting, namely, intrinsic motivation (Oudeyer & Kaplan 2007; Barto 2013) and empowerment (Salge et al. 2014; Salge & Guckelsberger 2016). The motivation of these works seems to be mainly practical, e.g., to provide a good balance between exploration and exploitation of the environment, and to enable the reinforcement learning system to explore its environment and to develop certain competences irrespective of an external reward.

Steward's pertinent remark that an action "cannot be regarded merely as the inevitable consequences of what has gone before" (Steward 2012, 72). A rigid, inevitable consequence of what has happened is not an action. It is useful to go beyond Steward here and to distinguish short-term aspects of this non-rigidity, which we discuss under the heading of *flexibility*, and longer-term aspects connected with *adaptivity* (see Müller and Briegel 2018).

In the short term, while an agent is acting in some way (e.g., trying to get something that it wants), the unpredictability of the environment will often provide challenges or obstacles. It is rare for an action to go through exactly as one would have thought it would. There are distractions, interferences, literal or metaphorical stumbling blocks, or even breakdowns (as when a piece of chalk breaks while you are writing on a blackboard, or when your phone runs out of battery). Agents need to be able to react in flexible ways to be able to continue to pursue their goals in the real world. As we said in discussing programming as control (Sect. 2.4.1.1), inflexible behaviour is a defeater of agency attribution. Flexibility builds on the availability of options for behaviour that are already presupposed by taking agents to be settlers of matters concerning their bodies, but it goes beyond that. Flexibility in the sense that interests us here means, in addition, that an agent can make sensible use of these various options such that it can hold on to a current goal. To give an everyday illustration, think about walking somewhere on a natural, uneven path. There are so many tiny things to adjust to, which we normally do competently and, as we would say, automatically: to avoid slippery bits and loose rocks, to place the feet more or less firmly on the ground and to adapt the pace to what is safe and useful in the circumstances. We and other animals can do this: we can hold on to our goal—in the example: to where we are trying to get—while making these necessary adjustments (which, depending on the circumstances, may also involve retracing our steps to try a different route). We do not normally have to let the fine details of a path dictate us where to go; we have a repertoire of skills that we can draw upon in the interest of our goals. Of course, there are no guarantees that we really have the flexibility required in a specific situation, and accidents happen, but they happen rarely, because we are good at moving around. (And of course, from an evolutionary perspective, we would not be around if we weren't.) So, non-rigidity of agency in the short term implies the meaningful and flexible use of open possibilities in the interest of our goals.

In the longer term, non-rigidity of behaviour also provides the basis for learning and adaptation. We are convinced that agents need to be learners, so that beings who do not build up a memory, or who do not adapt their behaviour based on their past experiences (including whether they got what they were trying to get), cannot be agents. Above we used susceptibility to training (as a form of learning) to distinguish between mere causal vs. agentic attribution (see Sect. 2.3.3). Agents can make use of their history, of what they have experienced, building up a repertoire of useful behaviours over time. This aspect of agency is very much underdeveloped

in the agency literature,⁸⁴ but it is absolutely crucial. All animal agents develop from an early to a mature stage. This process is, on the one hand, one of growth, which is a biological necessity for larger animals: for them it is simply not possible to have enough matter as material around to constitute a full-sized individual at birth or upon hatching. Growth in this sense may not be a necessity for artificial agents, who can be assembled by us in their mature size. But this process of development is, on the other hand, also one during which agents acquire various skills and behavioural routines that help them get along in the real world.⁸⁵ This may include learning to make sensible use of the possibilities for settling bodily matters that the agent has available in the first place. The process of learning to use one's limbs purposefully alone takes at least months in humans (Hafner et al. 2020; Nguyen et al. 2021). On top of that, agents develop skills that may be specific for the environment they find themselves in, including foraging or hunting strategies and a lot of skills needed for social interaction. Memorising past experiences, including what followed from trying to get certain things, agents thus build up their own learning history, which then becomes an important aspect of their individuality.

If the picture of agency sketched in terms of the above items (i)–(iv) is more or less correct, it follows that agency is a capacity that must have a general basis in the kind of agent in question, but also a specific basis in the concrete experiences and interactions with the environment that an agent builds up during its existence. (In many cases other agents will be an integral part of that environment.) Our own agency modelling will be strongly influenced by this focus on the diachronic, developmental aspects of agency.

Before we round off this chapter with a discussion of agency in science (Sect. 2.7), we briefly discuss three more important aspects of agency: agency and freedom (Sect. 2.6.2), the social nature of agency (Sect. 2.6.3), and our approach to agency modelling as projective simulation (Sect. 2.7.1).

2.6.2 Agency as Free Agency

Agency, as we said, involves specific capabilities of settling matters concerning the agent's body, and this is only possible for beings whose body is organised in such

⁸⁴ E.g., in the up-to-date *Handbook of Philosophy of Agency* (Ferrero 2022), remarks about learning occur in various places, but there is no discussion of learning and its relation to agency as a separate topic.

⁸⁵ This may provide another handle for making the distinction between animal agents vs. animals that cannot act: those that can must undergo some development and learning; the others need not. Bacteria, for example, start to exist more or less fully formed, while for humans that formation as an agent takes years. It is plausible that the formation as a moral agent takes even longer. Such a long time of development is, among other things, reflected in legal rules according to which different degrees of responsibility for one's actions are assumed at different ages. Full responsibility is normally reserved for adults.

a way that it leaves enough room, enough freedom for meaningful possibilities to settle. In this sense, then, agency is also a power to choose, and agency is fundamentally *free agency*. In the metaphysical sense here at issue, the “free” in “free agency” is merely explicative, not a substantive addition, and only unpacks what is contained in the notion of agency to begin with.

Freedom is, of course, a complicated business, and a few remarks are certainly in order. First, we will not enter a discussion of *free will* here, at least in the sense in which this notion is normally used in philosophy. Free will standardly denotes a capacity of *moral agents*, namely, that which makes them suitable targets for moral praise or blame. As we already said, we will not discuss moral agency here. The only remark we would like to make is that insofar as moral agency is a specific form of agency, free will must presuppose agency and all that comes with it. No agency, no free will.

Secondly, to claim that agency is free agency prompts the commonsensical response that surely, agents can be unfree, can't they? Oh yes they can, of course. For example, we all know that freedom of action, freedom of movement, freedom of speech, and other important freedoms are under threat in many parts of the world. Furthermore, some agents are locked up (for good or bad reasons) or even tied up against their will (for which good reasons are surely exceedingly rare). But all of this really only makes sense to discuss with respect to beings that are fundamentally free agents. The set of options among which they can settle anything (e.g., concerning their body) can be severely constrained, often by other agents, and often unjustly so. But even the person who is tied up has some possibilities left to settle. Minimally, they can settle what to think about. That is cold comfort to be sure, and no excuse for any harm done. But even a person in chains remains, fundamentally qua ontological category, an agent, and that is, a free agent.⁸⁶ This case of constrained agency needs to be sharply distinguished from other cases in which an agent can be brought (or bring itself) into situations in which its agency is temporarily suspended. In order to describe such cases adequately, it is important to distinguish between two meanings of being an agent, or of having agentive powers: (1) the normal use according to which a being is an agent if it can act at least sometimes during its existence, and (2) the temporally restricted use of being *currently* in the possession of agentive powers. While one is asleep, one is not an agent-(2), but one remains an agent-(1).⁸⁷

Sleep is important and beneficial, as we know. There are, however, also other cases in which one's agency is suspended, and in which one is effectively tem-

⁸⁶ At least superficially, this verdict aligns with some aspects of Sartre's notion of freedom in his *Being and Nothingness* (Sartre 1943).

⁸⁷ Agentive powers can also be diminished to various degrees rather than being fully suspended, e.g., through the use of various drugs. Such impairment may afflict an agent's power to choose, but also its power to learn and develop. Note that certainly with respect to us humans, there may be borderline cases of agency-(2), in which it is not really clear whether an agent-(1) is currently in possession of her agency-(2). Similarly, in the animal kingdom, there are also borderline cases of agency-(1).

porarily downgraded to a mere living being.⁸⁸ There is a difference between being tied up vs. being drugged to be in a state in which one lacks control altogether—the latter *does* take one's agency away temporarily. Sometimes this is wanted, as in certain medical procedures, but unless one consents, this will almost always be morally wrong. So, having one's freedom taken away has different meanings, some of which correspond to the suspension of agency, while others are more adequately described as *constraining* rather than *suspending* agency. (Which is often morally wrong too, of course.)

We should at least briefly mention two other phenomena that are often brought up in discussions of free vs. unfree agency, namely, compulsion or addiction, on the one hand, and coercion, on the other hand. Cases of the former are often grossly misdescribed in the philosophical literature.⁸⁹ It appears to be very rare for humans not to have *any* power to choose in cases of compulsion and addiction. It is true, however, that the costs associated with pattern-breaking choices (e.g., the associated feeling of psychological unrest) can be exceedingly high. In this, such cases *are* indeed quite close to cases of coercion, in which the high costs are not due to the agent's psychology, but are instead imposed by another agent—while the coerced agent's free agency is *not* suspended. That is indeed what is so malicious about coercion. Your money or your life, well, just a choice you have to make. And *you* make that choice. Note that coercion is thus also relevantly different from making merely causal use of another agent's body. If your neighbour threatens to call the police unless you stop the music and you give in to the coercion and stop it, *you* stop it. On the other hand, if she walks over and pulls the plug, *she* does that, even if she pulls the plug using your own hand as a tool.

Before moving on, we mention yet another issue that often comes up in this area of agency and freedom, namely, Frankfurt cases.⁹⁰ In Frankfurt cases, there is (allegedly) moral responsibility without alternative possibilities. Again, the literature is huge and controversial. We tend to side with those who hold that such cases are ill-described or ineffective (Keil 2023). For our purposes, however, we can leave the whole issue by the side. Frankfurt cases specifically concern moral agency, not agency simpliciter.

⁸⁸ And there is, of course, death, which suspends both agency and life permanently, and which thereby terminates the agent's and the living being's existence, leaving only its matter behind. A sign of this is that the organisation of that matter then rapidly falls apart through decomposition and decay.

⁸⁹ See, in contrast, Pickard (2021) for a well-informed study with many relevant references.

⁹⁰ So-called not because they are set in the city of Frankfurt, but because of the philosopher, Harry Frankfurt, who invented them, the *locus classicus* being his article "Alternate possibilities and moral responsibility" (Frankfurt 1969).

2.6.3 *Agency as Necessarily Social?*

We have defended the claim that artificial agency is at least conceptually possible, which implies that agents need not be biological entities.⁹¹ A somewhat related question is whether agents need to be social entities, or whether solitary agency is conceptually possible. As with the question of life, that may depend in part on how one interprets the notion of a social entity. For many higher animals that reproduce sexually, some form of social organisation is a necessity, but life is complicated. On the one hand, self-fertilisation is a possibility in some species, and even if genetic material from two individuals is involved, the social interaction may be rather indirect and limited, as in some fish or cephalopods. Dialectically, a single instance of a biological agent that does not live socially would already establish the conceptual possibility of non-social agents, and possibilities for the individualistic organisation of artificial agents seem even wider, given that care for offspring and social learning during development may not be needed. On the other hand, any form of communication involves a sender and a receiver and therefore implies some (limited) form of social interaction. So, agents who communicate have to be social to some extent.

That extent may, however, be quite limited. Social agency appears to be a graded notion, ranging from minimal forms of material message-passing that enable some form of collective agency, e.g., via pheromones, to discourse geared towards mutual understanding and on to highly organised and regulated forms of cooperative or joint agency, as in human team sports, business ventures, or governments.

In truly social cooperation (for example, in a dialogue), the agents involved (in that case, sender and receiver) need to recognise each other *as agents* in order to succeed. But even non-social agents (if there are any) will do well to recognise other agents as such; we already commented on the evolutionary benefits of this capacity at the beginning of Sect. 2.3.2. Such a capacity may thus be a practical necessity for biological agents. We leave it open, however, whether such a capacity is a conceptual necessity for agents generally speaking, i.e., also for artificial ones.

Overall, we refrain from passing a firm verdict on whether agency quite generally needs to be social. We are, however, convinced that *scientific agency* is social of necessity: Science must be a collective enterprise. Some forms of curious exploration and proto-science may be possible for a solitary individual, but real science needs intersubjective criticism as well as agreement to develop successfully (Oreskes 2019).

⁹¹ This does not fully settle the conceptual relation between agency and life. The fact that agency based on non-biological matter is conceptually possible, which we defended in Sect. 2.5 above, does not yet settle the question whether agents need to be *alive*. If one believes that artificial life is possible, one might be able to argue for that connection and still hold that agents need not be biological entities with respect to their material basis. As we said, there are reasons to be sceptical of artificial life (see note 74). Anyway, that issue is tangential to our discussion here.

2.7 Agency and Science

We started this Chapter with an exposition of the philosophical problem of agency, which we described as primarily a problem of ontology, of what there is. Being realists about agency, we affirm that agency has a foothold in our natural world, the world described by science and subject to scientific laws. Scientifically well studied entities such as the sun help us to anchor the notion of ontological emergence that we need for our explication of agency. That explication is based on a reflection on several of our practices involving agency, including agency attribution, and leaves open the possibility of artificial agency, i.e., the agency of technological artefacts based on current or future science.

In this final section of the Chapter we reflect on the interaction between agency and science, considering both directions. On the one hand, we consider the science of agency, or rather, a scientific approach to a more detailed understanding of the philosophical problem of agency: In Sect. 2.7.1, we briefly introduce the specific approach to agency modelling via Projective Simulation that will form the topic of Chap. 5 and, from a phenomenological point of view, of Chap. 6. On the other hand, we consider the role of agency in science: In Sect. 2.7.2, we briefly describe the common prejudice of considering science as pure theory, which informs much of philosophy of science, and set it off from experimental practice. Furthermore, we indicate in which way quantum science, already via the theory of quantum mechanics, makes it necessary to consider the agent behind the science for basic conceptual reasons. This discussion thus introduces the topics of Chaps. 3 and 4.

2.7.1 *A Target for Agency Modelling: Projective Simulation*

Science is a human activity, and it is an activity that is specifically human—a deeply social endeavour that depends on the coordinated interaction of many different human agents. In addition to human-human interaction, the scientists' use of and interrelation with more and more advanced technology is increasing in all sciences. Technology and science have long co-evolved, from instruments for observation (think of the telescope) to tools for manipulating nature at scales of magnitude and to degrees of precision unattainable by human beings alone (think of space probes reaching the limits of our solar system, or ion traps capturing single atoms) and on to ever more advanced tools for data analysis, experimental design, and automated experimental control. Technological developments have often enabled novel experiments, and experimental discoveries have often led to novel technologies. Rather than a clean separation of theory-centred pure science and application-oriented technology, there is really a continuum of techno-scientific practices (Russo 2022). Given the current development of technology, especially in AI, novel devices may soon play roles usually taken up by human agents. Indeed, if artificial agency turns out to be conceptually and technologically possible, it will probably be in some

fields of science that it will find some of its earliest applications, and advanced quantum technologies (e.g., the adaptive control needed to build and stabilise a working quantum computer) indeed suggest themselves as exemplary test cases. We will discuss the path towards such applications in Sect. 5.6. This is another reason why we find it useful to bring agency and quantum physics together in this book.

What about the science of agency? Given the features of agency that we have explicated in this chapter, how should we approach agency modelling? It is quite common to use models that compute an optimal action or plan to be executed. As we just argued, however, an agent needs to show sensible flexibility and adaptability, over short and longer time-scales, respectively. It has to have enough causal slack, enough open possibilities, to realise these features in a way that makes sense to it. Algorithmic optimisation may be a useful approach in “small world” scenarios such as games, but it does not provide useful guidance for agents in an open and changing environment. This means that agency modelling needs to be indeterministic at root. Judging by its basic features, an adaptive random walk is therefore a better model of an agent than a deterministic expected utility maximising algorithm.

Given the causal slack and the openness that comes from an indeterministic approach, however, the agent also has to make sensible use of it. Here we can connect again with Perl’s ladder of causation, whose third rung, “counterfactual thinking”, we linked with the third rung of agency attribution on our ladder of attribution, and ontologically, with real agency. The idea of counterfactual thinking is that an agent should, at least in some minimal sense, be responsive to considerations of “what would happen *if*”. The agent has to be able, at least sometimes, to react to counterfactual scenarios without having to fully try them out (which could be dangerous or otherwise costly). This is where the notion of *simulation* becomes useful. Via simulation, an agent can try out possible actions without trying them out in the real world; it can thereby protect its agential exploration of options from possibly excessive costs. Simulations come in variants of vastly different sophistication. We human agents can deal with explicit simulations, which inform us about expected outcomes of given scenarios. A government can, for example, run model-based simulations of certain economical measures and base their decisions on the outcome. But simulation can also mean something less explicit, and that sense of simulation is, in our view, important as a general background for agency modelling. It need not be that agents have explicit representations of simulation results *and then* use those to decide. Rather, a certain form of simulation can be part of the very process of decision finding. Simulation in that minimal sense of associating what is similar can integrate the virtual exploration of options with option selection in a dynamics that is intimately tied up with the agent’s learning history: the simulation dynamics can be shaped by past experiences, and it can thereby provide novel instances for learning through feedback. As we will discuss in more detail in Chaps. 5 and 6, such a simulation will thereby have a projective structure: the simulation will be the result of the agent’s projecting past experiences onto the current situation, and thereby, it will also be a means for the agent to project itself into the future, making sensible use of the options that its current environment affords.

These considerations point to the idea of agency modelling as Projective Simulation—a notion that prominently figures in the title of this book. In Chap. 5 we discuss Projective Simulation in formal detail as a highly promising indeterministic minimal agency model, providing our take on the basic science of agency. In the next two Chaps. 3 and 4, however, we first consider agency in science, and especially in quantum science.

2.7.2 *Agency in (Quantum) Science*

The philosophical reflection of the role of agency in science would seem to be a core task for philosophy of science. That sub-field of philosophy has, however, been mostly theory-centric, which is why the role of agency in science is under-researched. We will discuss this state of things briefly in Sect. 2.7.2.1. In quantum mechanics, however, the role of agency has forced itself on the agenda via a number of conceptual issues that go to the heart of that theory. We close this Chapter by a brief discussion of agency in quantum physics (Sect. 2.7.2.2).

2.7.2.1 The Absence of Agency in Philosophy of Science

Philosophically, science has for a long time been viewed primarily as a theoretical enterprise, and the separation of science and technology mentioned above lives on in the surprisingly stable separation of the philosophical fields of philosophy of science and philosophy of technology (Russo 2022, Ch. 2). Philosophy of science grew out of epistemology, the theory of knowledge, not out of action theory. Historically, its focus has primarily been on the structure of scientific theories, their logical properties, and their interrelations. A more recent focus on models brings philosophy of science in closer contact with scientific practice, but the picture remains very much theoretical. This state of things was diagnosed—and criticised—in a classic study by Hacking (1983), who notes that in many fields, there is a preference for theory right down to practices of distributing scientific credit. In the over forty years since, the practical aspects of doing science have been investigated in much more detail, but it remains a fact that practice-focused fields like the philosophy of experimentation are still underdeveloped. The main role of experiment is still seen as facilitating observations of nature, which can then be compared with the predictions of theories, confirming or falsifying them as the case may be. The lab is thereby divided up in a typical “front vs. back” dichotomy, where only the front, the presentation of the polished experimental results, is admitted to interact with theory.⁹² The lab is, however, not just a place where numbers, graphs, and tables of results are produced, but also part of the daily lives of those who

⁹² For the “front vs. back” dichotomy, see Goffman (1956). See also Hersh (1991).

work there. And even though experiments are mostly guided by theory or at least by theoretical hypotheses, the lab is not a mere outgrowth of scientific theorising. First and foremost, the lab is a place in which human beings interact with nature in a way that changes what would otherwise have been the case. Experimental practice consists to an important part of planned and controlled, meaningful human actions. The material world affords possibilities for us to intervene in various ways, and it takes effort and skill to intervene in the right way. The results of these interventions can then be observed and fed into the further scientific process.

Of course, there are sciences that have a strong focus on observation rather than on experiment, and which therefore seem far removed from human agency. Astronomy, on which much of modern science was modelled because of its astonishing precision, is a case in point—there is nothing we can do to influence the orbits of the planets and their moons, let alone the stars; all we can do there is observe, perfect the instruments and technologies of observation, and use our ever more precise observations in order to improve our models and theories. Another case in point is biological taxonomy, where the detailed observation and description of species and their characteristic traits is a key objective. Both these scientific practices are ancient. In the case of astronomy, the practice even predates Greek philosophy; and Aristotle was, among much else, a keen biological taxonomist.

Both practices, with their theoretical-observational focus, may have helped establish a view of the world as unchanging and deterministic, like clockwork; we will discuss that view in Sect. 3.4.2.1. Anscombe (1971, 143) has called this development “an intellectual disaster” and lamented that its grip on us is still strong. Indeed, we should by now know better. The scientific revolution of the 17th century was mainly driven not by advances in observation, but by a turn to experiment. The active role of the scientists was key here: modern science can only be done by real agents, because the notion of an experiment is grounded in real active intervention.⁹³ If anything, the importance of socially coordinated human agency in science has meanwhile only increased. The scale and scope of scientific experimentation, both in basic and in applied research and in industry, continues to grow. Just think of CERN, the European Organisation for Nuclear Research, which employs thousands of researchers and technicians and has an enormous budget of over a billion Euros per year. In our view, the size and the growing importance of experiments and the increasing role of automation in experimentation provides additional reasons for a

⁹³ This is not to say, of course, that developments of observational techniques have not been important. Such developments are ongoing, for example, in astrophysics, where they greatly enhance our understanding of nature. Accidental observations have sometimes also played a big role in science. And as mentioned in Sect. 1.4, there are so-called *natural experiments*, e.g., in epidemiology and in economics, in which what happens accidentally (or at least, what happens anyway) puts scientists in a position to draw conclusions very much like those that can be drawn from experiments. But in relation to real experiments, such cases are rare. And more fundamentally, one has to note that the concept of a natural experiment is parasitic upon the notion of a real experiment, as the epistemic standards of real experiments—based on real interventions—must be used to define what a natural experiment is.

detailed investigation of the role of agency in science. Quantum mechanics suggests itself as a particularly interesting case.

2.7.2.2 Agency in Quantum Mechanics

The active role of the experimenter is emphasised and indeed strengthened through a study of the most revolutionary fundamental physics of the early twentieth century, quantum mechanics. At least, this is how we view the matter, as we will lay down in more detail in Chap. 3. Briefly, the change with respect to experiment is this. In pre-quantum theories, the outcome of an experiment can (if all goes well) be understood as simply *revealing* what is out there in nature anyway, wholly independently of any experiment or observation. The perspectival considerations introduced by the special theory of relativity—e.g., the frame-relativity of simultaneity, which makes the question of whether two events are simultaneous or not depend on who is asking the question—do not change this picture fundamentally: through Minkowski space-time, there is still available a representation of what is *simply so* out there in nature, there to be read off from the mathematical structure. General relativity in addition incorporates gravitation, but still, the differential manifold that represents space-time in that theory provides a comprehensive account of all that can possibly be found out. In all these pre-quantum theories, the scientist who finds out about nature is, first and foremost, an *observer*. Experiments can be used to prepare certain configurations or states, which can then be studied with respect to their further properties or the dynamics of their evolution. That takes active intervention on the part of the experimenter, to be sure, but the result of such intervention is understood to be just an altered state of nature that is again fully determined with respect to all its properties—the activity of the experimenter fades to the background.⁹⁴

With the advent of quantum physics, this picture changes drastically. We will describe the theory of quantum mechanics in more detail in Chap. 3 and, with special emphasis on the role of agents, in Chap. 4. For now, the important point we wish to make is that theory is best understood as a theory not of what nature is like independently of all agency, but of what happens when we interact with nature in specific ways. In pre-quantum theories such as the ones mentioned above, the latter can be easily derived from the former, which is assumed to be given. But in quantum mechanics, no full picture of what is *simply so* out there in nature can be provided. The theory thus “only” captures what can happen in our interactions with nature. In this way, the notion of an agent—minimally, an entity that can put nature to the experimental test *in different ways* that are not settled by the system under investigation—enters the core of the theory.

⁹⁴ The overall coherence of this approach seems doubtful: On the one hand, it takes nature to be a deterministic “iron block” (to use the image made famous by William James), but on the other hand it still allows for some intervening powers within that block. It is unclear whether a robust a notion of intervention is really compatible with this approach.

Of course, as we know, saying all of this is controversial, and the literature on how to understand, or to interpret, quantum mechanics is immense, with fierce disagreements between some of the positions. We will describe a bit of that in later Chapters, but we make no claim of comprehensiveness. A key point that we can mention here already, however, is this. A complete characterisation of a quantum system at a specific time, given favourable circumstances, is to ascribe to it a (pure) quantum state $|\psi\rangle$. But even then one can ask different and often strictly incompatible questions about that system by doing different experiments. Again assuming favourable circumstances, such experiments are mathematically fully described by so-called “observables”, which from our active perspective is a misnomer, as the point is that choosing one rather than the other comes down to making an active difference in relation to the system. Anyway, these observables can be complementary in the mathematical sense that a determinate value with respect to one observable of a complementary pair rules out a determinate value with respect to the other. As a result of the mathematical structure of the theory, therefore, the quantum state $|\psi\rangle$ cannot be used to provide sharp values for the outcomes of all possible experiments that can be done on the system. Rather, the quantum state only provides *probabilities* for the different possible outcomes of possible experiments. And their realisation is contingent upon the choice of one experiment rather than the other. So it is that the notion of an experimenter’s active, free choice is incorporated into the core of the quantum formalism, as we will see in more detail in Chap. 3. Quantum mechanics thus sticks out as one area of science in which agency is fundamentally important.

References

- Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2015). *Molecular biology of the cell* (6th ed.). Garland Science.
- Anderson, P. W. (1972). More is different. Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047), 393–396.
- Anscombe, G. E. M. (1957). *Intention*. Blackwell. 2nd edition 1963, reprint Harvard University Press, Cambridge, MA, 2000.
- Anscombe, G. E. M. (1971). *Causality and determination*. Cambridge University Press. Page references are to the reprint in her *Collected Philosophical Papers*, Vol. 2, Oxford 1981, pp 133–147.
- Austin, J. L. (1956). A plea for excuses. *Proceedings of the Aristotelian Society*, 57, 1–30.
- Austin, J. L. (1962). *Sense and sensibilia*. Oxford University Press.
- Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In G. Baldassarre, & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems* (pp. 17–47). Springer.
- Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future: Agents and choices in our indeterminist world*. Oxford University Press.
- Bluvstein, D., Evered, S. J., Geim, A. A., Li, S. H., Zhou, H., Manovitz, T., Ebadi, S., Cain, M., Kalinowski, M., Hangleiter, D., Bonilla Ataides, J. P., Maskara, N., Cong, I., Gao, X., Sales Rodriguez, P., Karolyshyn, T., Semeghini, G., Gullans, M. J., Greiner, M., Vuletić, V., &

- Lukin, M. D. (2024). Logical quantum processor based on reconfigurable atom arrays. *Nature*, 626(7997), 58–65.
- Borexino Collaboration. (2018). Comprehensive measurement of pp-chain solar neutrinos. *Nature*, 562, 505–510.
- Bratman, M. E. (1987). *Intention, plans, and practical reason*. Harvard University Press.
- Brentano, F. (1874). *Psychologie vom empirischen Standpunkte*. Dunker und Humboldt. References are to the English translation, *Psychology from an Empirical Standpoint*, Routledge Classics ed., with a foreword by Tim Crane, London 2015.
- Briegel, H. J., & Müller, T. (2015). A chance for attributable agency. *Minds and Machines*, 25, 261–279.
- Bustamante, C. J., Chemla, Y. R., Liu, S., & Wang, M. D. (2021). Optical tweezers in single-molecule biophysics. *Nature Reviews Methods Primers*, 1(1), 25.
- Butterfield, J. (2011a). Emergence, reduction and supervenience: A varied landscape. *Foundations of Physics*, 41(6), 920–959.
- Butterfield, J. (2011b). Less is different: Emergence and reduction reconciled. *Foundations of Physics*, 41(6), 1065–1135.
- Carneiro, R. L. (2000). The transition from quantity to quality: A neglected causal mechanism in accounting for social evolution. *Proceedings of the National Academy of Sciences*, 97(23), 12926–12931.
- Cartwright, N. (1999). *The dappled world. A study of the boundaries of science*. Cambridge University Press.
- Chi, X., Li, Y., & Qiu, X. (2020). V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: Mechanism and regulation. *Immunology*, 160(3), 233–247.
- Cloeren, H. J. (1984). Ockham's razor. In J. Ritter & K. und Gründer (Eds.), *Historisches Wörterbuch der Philosophie*, Bd. 6. Schwabe Verlag. Sp. 1094–1095.
- Cyburt, R. H., Fields, B. D., Olive, K. A., & Yeh, T.-H. (2016). Big bang nucleosynthesis: Present status. *Reviews of Modern Physics*, 88, 015004.
- Davidson, D. (1978). Intending. *Philosophy of History and Action*, 11, 41–60.
- Del Colle, R. (2011). Miracles in Christianity. In G. H. Twelftree (Ed.) *The Cambridge companion to miracles* (pp. 235–253). Cambridge University Press.
- Dennett, D. C. (1984). *The intentional stance*. MIT Press.
- Dennett, D. C. (1991). *Consciousness explained*. Back Bay Books.
- Dennett, D. C. (2003). *Freedom evolves*. Vinking Books.
- Douguet, N., & Bartschat, K. (2018). Dynamics of tunneling ionization using Bohmian mechanics. *Physical Review A*, 97, 013402.
- Dowty, D. R. (1977). Toward a semantic analysis of verb aspect and the English 'imperfective' progressive. *Linguistics and Philosophy*, 1, 45–77.
- Driesch, H. (1914). *The history and theory of vitalism*. Macmillan.
- Dupré, J. (1995). *The disorder of things. Metaphysical foundations of the disunity of science*. Harvard University Press.
- Dupré, J. (2021). Causally powerful processes. *Synthese*, 199, 10667–10683.
- Dürr, D., Goldstein, S., Tumulka, R., & Zanghì, N. (2004). Bohmian mechanics and quantum field theory. *Physical Review Letters*, 93, 090402.
- Echeñique, J. (2012). *Aristotle's ethics and moral responsibility*. Cambridge University Press.
- Falkenburg, B., & Morrison, M. (Eds.). (2015). *Why more is different*. Springer.
- Ferrero, L. (Ed.). (2022). *The Routledge handbook of philosophy of agency*. Routledge.
- Fossa, F. (2021). Artificial agency and the game of semantic extension. *Interdisciplinary Science Reviews*, 46(4), 440–457.
- Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23), 829–839.
- Frege, G. (1893). *Grundgesetze der Arithmetik, Band I*. Verlag Hermann Pohle. English version, translated and edited by Philip A. Ebert and Marcus Rossberg, as Gottlob Frege: *Basic Laws of Arithmetic*, Oxford: Oxford University Press 2013.

- Gallow, J. D. (2022). The metaphysics of causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition. <https://plato.stanford.edu/archives/sum2022/entries/causation-metaphysics/>
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23, 121–123.
- Goffman, E. (1956). *The presentation of self in everyday life*. University of Edinburgh Social Sciences Research Centre.
- Goldstein, J. A. (2014). Anderson, P.W., “More is different. Broken symmetry and the nature of the hierarchical structure of science”, reprinted with an introduction. *Emergence: Complexity and Organization*, 16(3), 117–134.
- Gray, K., & Wegner, D. (2010). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, 14, 7–16.
- Gu, M., Weedbrook, C., Perales, A., & Nielsen, M. A. (2009). More really is different. *Physica D: Nonlinear Phenomena*, 238(9–10), 835–839.
- Hacking, I. (1983). *Representing and intervening*. Cambridge University Press.
- Hafner, V., Loviken, P., Pico Villalpando, A., & Schillaci, G. (2020). Prerequisites for an artificial self. *Frontiers in Neurobotics*, 14, 5.
- Hegel, G. F. W. (1812). *Wissenschaft der Logik*. Johann Leonhard Schrag. Translated by A. V. Miller as *Hegel's Science of Logic*, London: Allen and Unwin, 1969. References are to this translation.
- Heidegger, M. (1927). *Sein und Zeit*. Max Niemeyer. Quotations are from the English translation by John Macquarrie and Edward Robinson, *Being and Time*, Blackwell, Oxford 1962. Page references are to this translation, followed by the page number from the 19th German edition (Tübingen, Max Niemeyer Verlag, 2006) in brackets.
- Hersh, R. (1991). Mathematics has a front and a back. *Synthese*, 88, 127–133.
- Hofmann, C., Bray, A., Koch, W., Ni, H., & Shvetsov-Shilovski, N. I. (2021). Quantum battles in attoscience: tunnelling. *The European Physical Journal D*, 75(7), 208.
- Howard, D. (2007). Reduction and emergence in the physical sciences: Some lessons from the particle physics and condensed matter debate. In N. Murphy & W. R. Stoeger (Eds.), *Evolution and emergence: Systems, organisms, persons* (pp. 141–157). Oxford University Press.
- Humberstone, L. (1992). Direction of fit. *Mind*, 101(401), 59–84.
- Hume, D. (1748). *An enquiry concerning human understanding*. A. Millar. References are to the edition by Tom L. Beauchamp, New York: Oxford University Press, 2000.
- Humphreys, P. (2015). More is different ... sometimes: Ising models, emergence, and undecidability. In B. Falkenburg & M. Morrison (Eds.), *Why more is different* (pp. 137–152). Springer.
- Humphreys, P. (2016). *Emergence: A philosophical account*. Oxford University Press.
- Ichikawa, J. J., & Steup, M. (2024). The analysis of knowledge. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition. <https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/>
- Keil, G. (2023). Why Frankfurtian all-in can'ts are irrelevant to free will. *Inquiry*, 66(1), 143–164.
- Kim, J. (1998). *Mind in a physical world*. MIT Press.
- Kuhn, T. S. (1962). *The structure of scientific revolutions* (4th ed.). University of Chicago Press. With an introduction by Ian Hacking, 2012.
- Ladyman, J., & Ross, D. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press. Co-authored with David Spurrett and John Collie.
- Landsman, K. (2017). *Foundations of quantum theory. From classical concepts to operator algebras*. Springer.
- LeBlanc, F. (2010). *An introduction to stellar astrophysics*. Wiley.
- Levine, M. P. (2011). Philosophers on miracles. In G. H. Twelftree (Ed.), *The Cambridge companion to miracles* (pp. 291–308). Cambridge University Press.
- Lewes, G. H. (1875). *Problems of life and mind* (vol. 2). Kegan Paul, Trench, Turbner.
- Lewis, D. K. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74(4), 549–567.
- Löwe, B., & Müller, T. (2011). Data and phenomena in conceptual modelling. *Synthese*, 182(1), 131–148.

- Ma-Kellams, C. (2015). When perceiving the supernatural changes the natural: Religion and agency detection. *Journal of Cognition and Culture*, 15(3–4), 337–343.
- Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. Penguin.
- Marov, M. Y. (2015). The structure of the universe. In M. Y. Marov (Ed.), *The fundamentals of modern astrophysics: A survey of the cosmos from the home planet to space frontiers* (pp. 279–294). Springer.
- Maudlin, T. (2019). *Philosophy of physics. Quantum theory*. Princeton University Press.
- McGrew, T., & Larmer, R. (2024). Miracles. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition. <https://plato.stanford.edu/archives/sum2024/entries/miracles/>
- Meincke, A. S. (2024). Emergent properties. In A. R. J. Fisher & A.-S. Maurin (Eds.), *Routledge handbook of properties* (pp. 347–357). Routledge.
- Mitchell, K. J. (2023). *Free agents. How evolution gave us free will*. Princeton University Press.
- Moretti, L., & Piazza, T. (2018). Defeaters in current epistemology: Introduction to the special issue. *Synthese*, 195, 2845–2854.
- Mulder, J., Müller, T., Ometto, D., & Van Miltenburg, N. (2022). Causality and determination, powers and agency: Anscombean perspectives. *Synthese*, 200:452. Introduction to a topical collection consisting of 12 articles.
- Mulder, J. M. (2016). A vital challenge to materialism. *Philosophy*, 91, 153–182.
- Mulder, J. M. (2021). ‘Animals run about the world in all sorts of paths’: Varieties of indeterminism. *Synthese*, 199, 11945–11961.
- Müller, T. (2023). Taming Pereboom’s wild coincidences. *Mind*, 132(527), 789–802.
- Müller, T., & Briegel, H. J. (2018). A stochastic process model for free agency under indeterminism. *Dialectica*, 72(2), 219–252.
- Müller, T., & Placek, T. (2018). Defining determinism. *British Journal for the Philosophy of Science*, 69(1), 215–252.
- Nagel, T. (1986). *The view from nowhere*. Oxford University Press.
- Neuman, K. C., & Nagy, A. (2008). Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature Methods*, 5(6), 491–505.
- Nguyen, P. D. H., Georgie, Y. K., Kayhan, E., Eppl, M., Hafner, V. V., & Wermter, S. (2021). Sensorimotor representation learning for an “active self” in robots: A model survey. *Künstliche Intelligenz*, 32, 9–35.
- Nicholson, D. J., & Dupré, J. (Eds.) (2018). *Everything flows: Towards a processual philosophy of biology*. Oxford University Press.
- Noble, R., & Noble, D. (2021). Can reasons and values influence action: How might intentional agency work physiologically? *Journal for General Philosophy of Science*, 52(2), 277–295.
- O’Connor, T. (2021). Emergent properties. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition. <https://plato.stanford.edu/archives/win2021/entries/properties-emergent/>
- Odegard, V. H., & Schatz, D. G. (2006). Targeting of somatic hypermutation. *Nature Reviews Immunology*, 6(8), 573–583.
- Øhrstrøm, P., & Hasle, P. (2024). Future contingents. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition. <https://plato.stanford.edu/archives/fall2024/entries/future-contingents/>
- Øhrstrøm, P., & Jakobsen, D. (2022). Ockhamistic inspiration in modern tense-logic. In A. Santelli (Ed.), *Ockhamism and philosophy of time: Semantic and metaphysical issues concerning future contingents* (pp. 33–54). Springer.
- Oreskes, N. (2019). *Why trust science?* Princeton University Press.
- Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurobotics*, 1, 6.
- Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.
- Pereboom, D. (2001). *Living without free will*. Cambridge University Press.
- Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.

- Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference. Foundations and learning algorithms*. MIT Press.
- Pickard, H. (2021). Addiction and the self. *Noûs*, 55, 737–761.
- Pombo, O. (2011). Neurath and the encyclopaedic project of unity of science. In J. Symons, O. Pombo, & J. M. Torres (Eds.), *Otto Neurath and the unity of science* (pp. 59–70). Springer.
- Quine, W. V. O. (1948). On what there is. *Review of Metaphysics*, 2(5), 21–38.
- Robb, D., Heil, J., & Gibb, S. (2023). Mental causation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition. <https://plato.stanford.edu/archives/spr2023/entries/mental-causation/>
- Russell, B. (1912). On the notion of cause. *Proceedings of the Aristotelian Society*, 13, 1–26.
- Russell, B. (1948). *Human knowledge: Its scope and limits*. Allen and Unwin.
- Russell, S., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education.
- Russo, F. (2022). *Techno-scientific practices. An informational approach*. Rowman & Littlefield.
- Salge, C., Glackin, C., & Polani, D. (2014). Empowerment – An introduction. In M. Prokopenko (Ed.), *Guided self-organization: Inception* (pp. 67–114). Springer.
- Salge, C., & Guckelsberger, C. (2016). Does empowerment maximisation allow for enactive artificial agents? In *Proceedings of the Artificial Life Conference 2016* (pp. 704–711). MIT Press.
- Sartenaer, O., & Baysan, U. (2021). Non-standard approaches to emergence: Introduction to the special issue. *Synthese*, 199, 7773–7776.
- Sartre, J.-P. (1943). *L'Être et le néant*. Gallimard. English translation by Sarah Richmond: *Being and Nothingness*, Routledge, London 2018.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140–1144.
- Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
- Steward, H. (2020). Free will and external reality: Two scepticisms compared. *Proceedings of the Aristotelian Society*, 120(1), 1–20.
- Steward, H. (2021). What is determinism? Why we should ditch the entailment definition. In M. Hausmann & J. Noller (Eds.), *Free will: Historical and analytic perspectives* (pp. 17–43). Springer.
- Steward, H. (2022). Animal agency. In L. Ferrero (Ed.), *The Routledge handbook of philosophy of agency* (pp. 101–108). Routledge.
- Straub, I. (2016). 'It looks like a human!' The interrelation of social presence, interaction and agency ascription: A case study about the effects of an android robot on social agency ascription. *AI & Society*, 31(4), 553–571.
- Strawson, P. F. (1959). *Individuals. An essay in descriptive metaphysics*. Routledge.
- Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 187–211.
- Strawson, P. F. (1992). *Analysis and metaphysics*. Oxford University Press.
- Tennant, N. (2023). Logicism and neologicism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition. <https://plato.stanford.edu/archives/win2023/entries/logicism/>
- Thomason, R. H., & Horty, J. (2022). Artificial and machine agency. In L. Ferrero (Ed.), *The Routledge handbook of philosophy of agency* (pp. 366–375). Routledge.
- Thompson, M. (2008). *Life and action*. Harvard University Press.
- t'Hooft, G. (2016). *The cellular automaton interpretation of quantum mechanics*. Springer.
- Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.
- Tumulka, R. (2021). Bohmian mechanics. In E. Knox & A. Wilson (Eds.), *Routledge companion to the philosophy of physics* (chap. 17). Routledge.
- Uebel, T. E. (2007). *Empiricism at the crossroads: The Vienna Circle's protocol-sentence debate*. Open Court.

- Vaidman, L. (2021). Many-worlds interpretation of quantum mechanics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition. <https://plato.stanford.edu/archives/fall2021/entries/qm-manyworlds/>
- Van Inwagen, P. (1983). *An essay on free will*. Oxford University Press.
- Van Lier, M. (2025). *Modeling the artificial agent*. Ph.D. Thesis, University of Konstanz.
- Van Riel, R., & Van Gulick, R. (2025). Scientific reduction. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2025 edition. <https://plato.stanford.edu/archives/sum2025/entries/scientific-reduction/>
- Velleman, J. D. (1992). What happens when someone acts? *Mind*, 101(403), 461–481.
- Vladimirov, N., & Sourjik, V. (2009). Chemotaxis: How bacteria use memory. *Biological Chemistry*, 390(11), 1097–1104.
- Watson, G. (1987). Responsibility and the limits of evil: Variations on a Strawsonian theme. In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility* (pp. 119–148). Cornell University Press.
- Wegner, D. (2002). *The illusion of conscious will*. MIT Press.
- Wellmann, J. (2024). *Biological motion*. Zone Books.
- Wu, T., & Tegmark, M. (2019). Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E*, 100, 033311.
- Yates, D. (2016). Demystifying emergence. *Ergo*, 3, 809–841.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 3

Quantum Mechanics



Quantum mechanics (QM) is a highly successful theory that has a host of important technological applications. Virtually all of modern electronics, from digital cameras to computers to LED light bulbs, is based on the successful harnessing of quantum phenomena. Still, since its inception in the 1920s, QM has also been a source of numerous conceptual and philosophical puzzles, and many of these puzzles are still being discussed today. Before this background, bringing together QM and agency—another source of conceptual and philosophical puzzles, as we just saw in Chap. 2—will surely raise some eyebrows. The danger, which we are well aware of, is to employ the method of *obscurum per obscuris*, to explain what one does not understand by relating it to what one understands even less. Still, we want to advance and defend a general claim about the interrelation of agency and QM, in the spirit of our book’s subtitle to provide “quantum-mechanical perspectives on agency”: QM and agency are virtually made for each other. QM is a theory for and by agents interacting with nature, and unless one subscribes to agency scepticism, QM describes a world in which the future is open. QM thereby offers resources that support our realist view of ontological levels, up to agency, the most important resource being fundamental indeterminism. Before we address these claims in Sect. 3.6, we provide some background on the theory and its main mathematical and conceptual features. Our emphasis is on the conceptual side, but mathematical details are given where that seems important for concreteness and precision.

3.1 Quantum Mechanics: A Brief Historical Introduction

Quantum mechanics has long been considered a philosophical and conceptual conundrum, and its fundamental notions are a source of ongoing controversy.¹

¹ For references, see, e.g., Healey (2023) and Myrvold (2022).

There is disagreement even about what QM's fundamental notions *are* to begin with. Quite generally, no theory, no worldview can exist without some accepted foundations that are not explained, but simply accepted. Not everything can be defined or derived. That masses and forces exist is not a discovery of Newtonian mechanics, but part of the fundamental framework that makes this theory possible. Similarly, classical electrodynamics simply accepts that there are electric charges; it goes on to explain how they behave, but it does not have any account for why they exist. (Much less does it have an account for why charge is quantised and why all electrons have *exactly* the same charge.) Quantum mechanics does not live in a conceptual void either. But the question of what the conceptual foundations of that theory are has been fiercely debated since its early days.² We will comment on a number of conceptual issues in what follows. It seems fitting, however, to start with a brief historical introduction.

3.1.1 *Physics Around 1900*

Around 1900, large parts of physics were making very good progress at what, in the language of Kuhn (1962), we can call *normal science*.³ There were, however, also a number of recalcitrant phenomena, mostly to do with discontinuities in nature, that could not be adequately explained. Striking examples of such discontinuous phenomena were the line spectra of the chemical elements, which had been observed by prismatic diffraction of light from coloured flames since the 1850s; another riddle was the photoelectric effect, in which the frequency (rather than the intensity) of absorbed light was shown to have a discontinuous, all-or-nothing effect on the current between metal plates at different electrical potential. In both cases, early aspects of the emerging quantum theory, viz., the quantisation of energy in the form of so-called light quanta or photons (fixed packets of an energy linearly related to the light's frequency), helped explain the phenomena. Incidentally, the hypothesis of quantised energy had at around that time also been found useful for the explanation of another conundrum, the (continuous) spectrum of blackbody radiation.⁴

² The literature on quantum foundations is enormous, and while we will provide some references along the way, we cannot attempt to give a balanced overview. Englert (2013) provides a useful, opinionated discussion of conceptual problems of QM and argues that the theory is best viewed as being built on the acceptance of random events. While our approach focuses more strongly on agency, our view has much in common with his.

³ There are anecdotes suggesting that promising young people around 1900 were told to stay away from physics, as that science was supposedly nearing completion. Some quotes, e.g., Lord Kelvin's alleged prediction of the end of physics ca. 1910, are dubious. But at any rate, physics Nobel laureate Michelson claimed in a lecture given in 1899 that "[t]he more important fundamental laws and facts of physical science have all been discovered, and these are now so firmly established that the possibility of their ever being supplanted in consequence of new discoveries is exceedingly remote" (Michelson, 1903, 24).

⁴ In Kuhnian terms, the line spectra, the photoelectric effect, and blackbody radiation were thus *anomalies* of the relevant prevailing theory of the time, electrodynamics.

A first conceptual novelty of the emerging quantum theory was, accordingly, its rejection of the ancient principle that there are no leaps in nature (“*natura non facit saltus*”); thus the “quantum” in QM. Light, which in the nineteenth century was thought to be a pure wave phenomenon, was now hypothesised to have corpuscular aspects; similarly, later it also became plausible to assume that matter has wave-like aspects. This is the source of one much-discussed aspect of QM, the so-called “wave-particle duality”. Apparently, there are two different and incompatible ways of describing phenomena like a beam of light (or of electrons): as a wave or as a stream of corpuscles.

Such dualities are not uncommon, and might just be an effect of taking different perspectives. It is, after all, well known that taking different perspectives on a problem can make a huge difference to understanding it, with one perspective effectively hiding the solution and the other offering the solution on a silver platter.⁵ Furthermore, the importance of different perspectives for physical theories was at the time being driven home by relativity theory, another revolutionary successor to electrodynamics, in which the perspective of an observer plays a crucial role, e.g., for the time ordering of events. So might it not just be that sometimes the wave perspective helps, and sometimes it is the particle perspective?

In QM, on the one hand, one might say that matters turned worse, as something akin to the assumption of incompatible “perspectives” on a system was found to be a crucial and irreducible aspect of the mathematical structure of the theory quite generally, not limited to wave-particle duality; we will discuss this aspect of QM under the heading of complementarity below. On the other hand, as we will argue, one might also say that in QM, matters turned out to be closer to what we are used to in everyday contexts of interacting with the world around us.

To be more specific, in Sect. 3.2 we give a brief overview of the mathematical framework of QM, which was established soon after the pioneering work of the mid-1920s, and which has remained the theory’s stable basis since then. We introduce that framework before the background of the astonishing Stern-Gerlach experiment, which was conducted in 1922 and which illustrates many of the counterintuitive features of QM in an especially clear way. We describe this experiment in Sect. 3.1.3. With a view to later discussions of modelling, however, we first describe in a little more detail the riddles and problems that confronted atomic physics in the early 1920s. A key riddle was posed by the so-called anomalous Zeeman effect.

⁵ Consider Black’s (1946, 157) problem to cover a chessboard, with two opposing corner fields removed, by dominoes each of which covers two adjacent fields. The helpful perspective here is to see that each domino has to cover one white and one black field, while the two corner fields that were removed are of the same colour. When one takes that perspective on the problem, one sees immediately that it cannot be solved, and that it is no use trying it out.

3.1.2 Atomic Spectra and the Anomalous Zeeman Effect

What was to become quantum mechanics was, in the early twentieth century, to a large extent atomic physics. Apart from a drive to understand the structure of the periodic table of the chemical elements (e.g., the length of the periods 2, 8, 18, 32: why the series $2n^2$?), this mostly meant the study of the behaviour of atoms via the radiation—visible light, but also X-rays—that they emit in various circumstances. Advances in experimental practices and in optics (e.g., Rowland diffraction gratings for high-resolution spectroscopy) made available a wealth of empirical data about atomic spectra. Many different chemical elements were studied, as was the influence of electric and magnetic fields on the spectral lines. Models on the basis of classical electrodynamics could not account for the discrete nature of these lines, but Bohr's 1913 model of the atom was celebrated as successfully explaining the spectrum of hydrogen: the atom's single electron was taken to orbit the nucleus at only discrete allowed distances, and the spectral lines then were accounted for as packets of energy (photons) emitted or absorbed when the electron transitioned between these orbits. Heavier chemical elements, starting already with helium, were, however, outside the model's scope. Another problem concerned the complex structure of the line spectra, which was known for the heavier elements to show a wealth of detail, especially under the influence of external fields.⁶ The most recalcitrant of these complexities was generally taken to be the so-called anomalous Zeeman effect.

Zeeman had in 1896 managed to study spectral lines in a strong magnetic field and shown a splitting of the lines. An explanation for a threefold splitting was found by Lorentz, lending support to the then still hypothetical idea that the atomic core was orbited by electrons that were causally involved in the production of the radiation.⁷ The threefold splitting was known as the “normal” Zeeman effect, even though it is actually a rather rare case—most splittings are of different multiplicity, and even numbers are in fact very common. Such cases were sometimes not taken seriously or somehow discussed away by attributing them to secondary effects, but this was not always possible. There remained the incontrovertible fact of the “anomalous” Zeeman effect, so-called not because it does not normally occur, but because it resisted attempts at explanation with the resources available up to 1924.⁸

At that time, different approaches to solving the riddle of even-numbered splittings of lines in atomic spectra were around—some partial models involving a

⁶ See Forman (1968) for details. As Forman notes, the idea that the hydrogen spectrum itself also had a complex structure was basically inconceivable for a while, despite at least some contrary evidence, and so the two problems of accounting for heavier elements and for complex spectra were taken to be fully aligned (*ibid.*, 164).

⁷ See Hentschel (1996) for some historical details, including an assessment of the formidable difficulty of carrying out the original experiment. Zeeman and Lorentz shared the 1902 Nobel prize in physics for their work.

⁸ See again Forman (1968) for details. See also Schmidt-Böcking et al. (2023) for details of Landé's work that, starting in the early 1920s, established an empirical description of the effect, although without providing an underlying model.

magnetic contribution of the atom's core, some involving various relativistic effects, some purely at the level of formulae that reproduced the observed multiplicity numerically. No comprehensive solution was, however, possible. As *we can now* say, that was because the different approaches did not contain a rich enough set of fundamental concepts to provide a successful explanation. The necessary addition was offered in a brief note by Uhlenbeck and Goudsmit (1925), two young physicists, who suggested that the *electron* could be the source of a twofold splitting via a novel internal degree of freedom, now called *spin*.⁹ That idea was subsequently shown to explain a large number of recalcitrant phenomena including hitherto troublesome aspects of X-ray spectra and of the fine structure of the hydrogen spectrum; it also provided the key for an explanation of the anomalous Zeeman effect as due to the interaction of the electron spin with the external magnetic field (Uhlenbeck & Goudsmit, 1926). Once some more details had been worked out (Thomas, 1926), the notion of spin was taken up by the community at large due to its enormous explanatory power. In fact, it also fitted nicely with Pauli's (1925) exclusion principle. That principle holds that no two electrons can coincide in the specification of their full set of degrees of freedom ("quantum numbers"). Pauli had to postulate an extra, non-classical two-valued degree of freedom in order to explain the structure of the periodic table and various other phenomena; spin turned out, after all, to furnish that degree of freedom (Pauli, 1927).

The introduction of spin came at the time when the mathematical structure of QM was successfully being established in two different ways by Heisenberg and Schrödinger. Before we proceed to introduce that mathematical structure (Sect. 3.2), we describe the Stern-Gerlach experiment, which, in hindsight, provides an additional motivation for the introduction of the concept of spin, even though the connection was made only many years after the experiment.

3.1.3 *The Stern-Gerlach Experiment*

The Stern-Gerlach experiment, named after Otto Stern and Walther Gerlach, was set up in an attempt to test and possibly falsify the hypothesis of a certain type of directional quantification caused by a magnetic field. Contrary to what at least one of the experimenters had expected, such a quantification, evidenced by a twofold splitting of the atoms in the field, was indeed found, but it remained impossible to satisfactorily explain the results for several years.

The experiment made use of the novel technology of producing beams of atoms in a high vacuum, and subjected such a beam to a specifically chosen magnetic field.

⁹ In his personal recollection, Goudsmit (1971) reports that their mentor Ehrenfest had said that it was a nice idea, and though it might be wrong, the young men did not yet have a reputation and therefore had nothing to lose. Goudsmit gives evidence that earlier attempts at introducing something like spin were ridiculed by Pauli to such an extent that they were left unpublished.— See Commins (2012) for a contemporary overview of electron spin and its history.

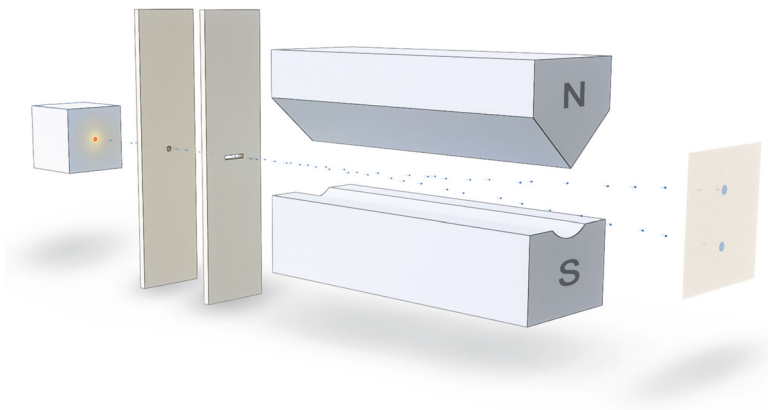


Fig. 3.1 The set-up of the 1922 Stern-Gerlach experiment. Silver atoms emanate from an oven (left). The atomic beam is collimated through two apertures, traverses the inhomogeneous magnetic field generated by an electromagnet (middle right) and then impinges on a screen (right)

Its main result was that the beam was split into *two* distinct components, while the classical expectation was that the beam should retain *one* continuous component and just be broadened while passing through the field. Several modern textbooks of QM use the experiment to illustrate a number of key basic features of QM in one uniform framework,¹⁰ and we will use it to the same end here. There are several advantages to this approach: The mathematical structure that describes the surprising aspects of the Stern-Gerlach experiment is much simpler than, say, the mathematics required to describe the fine structure of the spectrum of hydrogen, and the experiment brings out the already mentioned phenomenon of complementarity in an especially clear way. Furthermore, as we will see in Sect. 3.3.3 below, the Stern-Gerlach experiment also offers a helpful perspective on the role of irreversibility in quantum measurement, and it illustrates the notion of a qubit, a fundamental resource of quantum information theory, which we will discuss in Sects. 3.3.3.2 and 3.4.4.

Set-up and Results The experimental set-up is schematically pictured in Fig. 3.1. Silver is vaporised in an electrically heated oven. The silver vapour exits the oven through a small hole. The resulting beam of electrically neutral silver atoms, flying at a speed of the order of 1000 m/s in a high vacuum tube of ca. 10 cm length, is further collimated through two small apertures, and at the end of the tube the beam impinges on a glass plate, which accordingly records the position of the atoms at the end of their flight. In the field-free set-up, as expected, the distribution of silver on the plate is a geometrical image of the last slit. The key experimental intervention is to switch on a magnetic field perpendicular to the beam. That field is produced by an electromagnet whose one pole has a sharp edge, while the other

¹⁰ See, for example, Schwinger (2001), Englert (2006a), and Sakurai and Napolitano (2020).

pole is flat, resulting in an inhomogeneous field. It was known that the silver atoms have a magnetic dipole moment. Classically one expects that these dipoles should have random orientations when the atoms emerge from the oven, and thus the beam should be simply broadened by passing through the field, resulting in a smeared-out geometrical image of the slit. One finds, however, that the beam is split into two distinct, spatially separated components (shown on the right of Fig. 3.1). In the original experiment, this separation was around 0.1 mm, which is approximately equal to the thickness of the geometrical image of the slit in the field-free set-up. The separation appears to indicate that in the inhomogeneous field, the dipoles align in one of only two possible orientations, indicating a quantisation of the direction that must be of a novel type. At the time, no mechanism was known that could explain how this should come about.¹¹

As described in the previous section, during the course of the 1920s, detailed investigations of atomic spectra suggested that the electron must possess an internal degree of freedom, spin, that interacts with an external magnetic field in puzzling ways and which has no direct classical analogue. From our present perspective, the result of the Stern-Gerlach experiment can fairly easily be explained if one assumes that the silver atom's magnetic moment is due to its outermost electron, whose spin can, for any direction along which it is measured, assume just one of two possible values.¹² These possible values, namely, $\pm\frac{1}{2}\hbar = \pm\frac{1}{4\pi}h$, are related to Planck's constant $h \simeq 6.626 \cdot 10^{-34} Js$. In what follows, we will adopt the usual contemporary perspective of describing the experiment in terms of spin. Considering first a single Stern-Gerlach set-up and then several of them in sequence, one can illustrate some of the most puzzling aspects of QM.

Quantization of Measurement Results Along any Direction It is customary to describe the atomic beam as moving in the y -direction in a standard right-handed coordinate frame. The inhomogeneous field is initially oriented in the vertical, z -direction. Using Dirac's bra-ket formalism, which we will explain in Sect. 3.2.1, one can denote the two possible output states of the atoms after interaction with the inhomogeneous field in mnemonic form as kets $|\uparrow, z\rangle$, “up along z -direction”, and $|\downarrow, z\rangle$, “down along z -direction”, respectively.

As the environment of the experiment can be assumed to be rotationally symmetric around the direction of the beam, rotating the apparatus (thus, the magnet, and thus, the field) should not make a difference, and this is what one finds—the beam is split into two components along any chosen direction of the

¹¹ The original publication is Gerlach and Stern (1922). The puzzlement about the experiment is aptly expressed, e.g., by Einstein and Ehrenfest (1922). See Schmidt-Böcking et al. (2016) for an overview.

¹² While that explanation may seem obvious in hindsight, and some popular descriptions even suggest that the experiment was instrumental in establishing the notion of spin, that connection was not at all made at the time. According to Schmidt-Böcking et al. (2023, 11), the first explanation of the Stern-Gerlach splitting in terms of spin was only given in 1937. (The “1927” in the text is obviously a misprint, and is corrected in the related account in Friedrich (2023).)

inhomogeneous field perpendicular to the beam, e.g., also in x -direction.¹³ This indicates, as one also expects, that the atoms do not leave the oven with any preferred direction of alignment. Generally, for a unit vector $\vec{n}$ (here in the x - z -plane), one denotes the two resulting output states after interaction with an $\vec{n}$ -aligned apparatus as $|\uparrow, \vec{n}\rangle$ and $|\downarrow, \vec{n}\rangle$.¹⁴

The beam of atoms is very dim—in the original experiment, after 8 hours of operation, the silver spots on the glass plate were still so faint that they had to be chemically developed to enable the microscopic measurements of their separation.¹⁵ Ideally, one can assume that the beam is so dim that the atoms pass the apparatus one at a time, so that interaction between the individual atoms is excluded. In such a case, each atom must somehow independently of all others “decide” to be deflected in the positive or negative direction of the field. This process appears to be fundamentally probabilistic, but so far, one might also speculate that some inhomogeneity of the orientation of the atoms coming out of the oven has a decisive effect here. Further elucidation—and puzzlement—is provided by operating several Stern-Gerlach devices in sequence.

Sequential Stern-Gerlach Experiments Several interesting aspects of QM can be illustrated by considering what happens if one uses the output of one Stern-Gerlach experiment as the input of a consecutive one.¹⁶ The simplest such set-up consists of two apparatuses that are aligned in the same direction. Suppose that we take the upper output of the first apparatus (oriented along the z -direction), for which we write $|\uparrow, z\rangle$, and feed it into a second, parallel apparatus, also oriented along z (Fig. 3.2a). We know that this second apparatus can at most have two output beams, an upper ($\uparrow$) and a lower ($\downarrow$) one, like the first. The question is whether there will be any atoms coming out of the lower channel of the second apparatus, or whether the outcome of the first apparatus persists.

While initial experiments conducted in the 1930s were overall not fully conclusive,¹⁷ later experimental evidence shows clearly that in such an arrangement, the result of the first splitting is preserved: all atoms from the upper channel of the first apparatus also come out at the upper channel of the second one, and the lower channel of that second apparatus is empty. In this case, it makes sense to think of

¹³ The effect of gravity, which physically distinguishes the x - from the z -direction, can be neglected: during the time of flight of around 10^{-4} s, the atoms fall a distance of the order of 10^{-4} mm, a thousandth of the image width.

¹⁴ The above notation, $|\uparrow, z\rangle$, could thus more properly be written as $|\uparrow, \vec{e}_z\rangle$, where $\vec{e}_z$ is the unit vector pointing in the direction of the z -axis. We stick with the simplified notation, and we use an analogous simplification for x and y .

¹⁵ See Friedrich and Herschbach (2003) for the nice anecdote of how the initial idea of developing the silver deposits came from the effect of the sulphurous smoke of a bad cigar.

¹⁶ The experimental study of consecutive Stern-Gerlach devices was suggested by Einstein in letters dating from early 1928, see Schmidt-Böcking et al. (2016, 344f.) for details. In their paper from 1932 reporting preliminary results, Phipps and Stern (1932, 185) say that such an experiment had been proposed on numerous occasions (“wiederholt [...] vorgeschlagen”).

¹⁷ See Frisch and Segré (1933) as well as the analysis in Schmidt-Böcking et al. (2016).

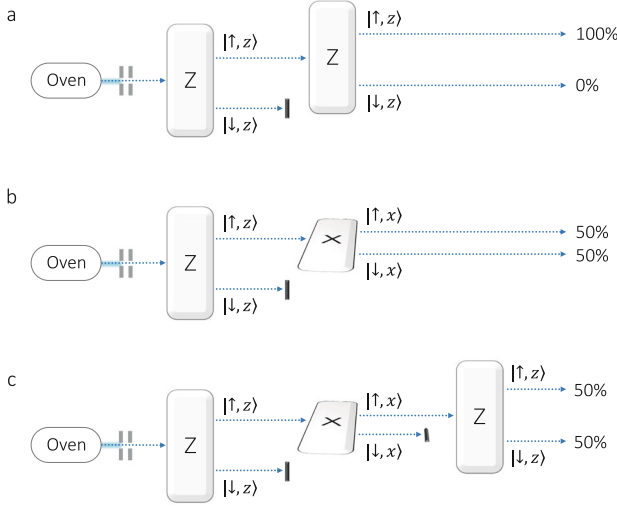


Fig. 3.2 Several options for sequential Stern-Gerlach experiments. See main text for details

the spin of each of the atoms coming out of the first apparatus as “aligned positively along the z -direction”. This is useful not just for an explanation of the sequential arrangement of two *parallel* Stern-Gerlach devices, but also for understanding the sequential arrangement of two Stern-Gerlach devices that are *not* aligned in parallel. We can turn the second apparatus in the x - z plane (the plane perpendicular to the beam) to some orientation $\vec{n}$ that is rotated away from the z -axis by an angle φ . Then, once $\varphi \neq 0$, there will be some intensity in the lower ($\downarrow$) channel of the second apparatus. One finds that the intensity of the upper beam at the output of the second apparatus (in units of the intensity of the incoming beam of atoms in the state $|\uparrow, z\rangle$) varies with φ as

$$\frac{1 + \cos \varphi}{2} = \cos^2(\varphi/2), \quad (3.1)$$

while the intensity of the lower beam is

$$\frac{1 - \cos \varphi}{2} = \sin^2(\varphi/2). \quad (3.2)$$

As we said, the experiment can be conducted with very dim beams, so that the beam intensity should rather be understood as a single-case probability for what happens to one single atom passing through the apparatus. That is, given an atom from the upper channel of the first apparatus, whose spin state is described by $|\uparrow, z\rangle$, the probability of that atom’s coming out of the upper ($\uparrow$) or lower ($\downarrow$) channel of a second apparatus aligned along direction $\vec{n}$, which is rotated by an angle φ with respect to the z -axis, is

$$\Pr(\uparrow, \vec{n} \mid \uparrow, z) = \cos^2(\varphi/2); \quad \Pr(\downarrow, \vec{n} \mid \uparrow, z) = \sin^2(\varphi/2). \quad (3.3)$$

In the extreme case of $\varphi = \pi$ (180° , a turn by a half circle), upper and lower channel along z are simply switched, and accordingly all the atoms will come out at the lower, $\downarrow$ -channel of the second apparatus. For the x -direction, which corresponds to $\varphi = \pi/2$ (a turn by 90°), both channels of the second apparatus will have the same intensity, so that there is an equal, 50:50 chance for each atom to come out in the upper or lower channel (Fig. 3.2b). This means that the atoms coming out of the first, z -oriented apparatus in the upper channel (spin state $|\uparrow, z\rangle$) have no bias one way or the other for a subsequent splitting along the x -direction. Their ascribed state $|\uparrow, z\rangle$ contains no specific information at all regarding a splitting into $|\uparrow, x\rangle$ or $|\downarrow, x\rangle$.

Rather than turn the second apparatus, which makes the experimental set-up quite challenging, one can also consider arrangements that include two *parallel* Stern-Gerlach experiments and an intervention via a magnetic field in between them.¹⁸ For example, one can shield off the influences of the inhomogeneous fields between the two devices and switch on a *homogeneous* magnetic field in the space between them. By tuning that field in different ways, one can—to use the classical image that was also found useful for our above considerations—rotate the direction of alignment of the spins (rather than the orientation of the second experiment) away from z by an angle φ . In this case, we also recover the quantitative relations from above, as we will discuss in detail in Sect. 3.3.3.¹⁹

Stern-Gerlach Orientations and Complementarity Bohr coined the term *complementarity* to describe a situation like the one that can be exhibited by two sequential Stern-Gerlach devices oriented along the z - and then the x -direction: full information about the system (in this case, a silver atom) with respect to one property (here, spin up or down along the z -direction) excludes any information with respect to another possible system property (here, spin up or down along the x -direction). These two possible system properties are complementary in the sense that they correspond to two different, incompatible questions that one may ask about the system. If one focuses on finding out about the z -orientation, there cannot be any information regarding the x -orientation, and vice versa. It is as if these ways of addressing the system amounted to incompatible perspectives on the system that cannot be combined into a single, coherent whole—which calls into question the whole idea of system properties.²⁰

¹⁸ In fact, Einstein's suggestion for an actual experiment, mentioned in Footnote 16, was along these lines.

¹⁹ See Eq. (3.80).—Luckily we do not have to keep track of sign conventions about the rotation, as the results are symmetrical under the exchange $\varphi \leftrightarrow -\varphi$.

²⁰ The complementarity relation in fact holds for any two directions of spin projection that make an angle of 90° , irrespective of the absolute directions. Based on this insight, one can already derive the commutation relations among the spin projection operators. See, e.g., Sakurai and Napolitano (2020, Ch. 3.2). See also our discussion in Sect. 3.3.3 below.

Further insights into the issue of complementarity are possible if one considers *three* sequential Stern-Gerlach devices.

Three Sequential Stern-Gerlach Devices Sequencing three Stern-Gerlach devices allows one to answer the following crucial question: Is any of the initial z -orientation retained in an atom that first passes a z -directed Stern-Gerlach apparatus and then a second, x -oriented one? For the first two devices, we thus consider the above case of turning the second apparatus by an angle of $\varphi = \pi/2$. The first apparatus, whose input consists of the atoms from the oven, outputs half of them in its upper channel (spin state $|\uparrow, z\rangle$). We block the output of the lower channel and feed the upper beam into the second, x -oriented apparatus. As we said, its two output channels then have the same intensity, and as we noted, this means that we have *probabilities for each single atom*, which enters the second apparatus in state $|\uparrow, z\rangle$, to come out spin up or down along the x -direction, according to Eq. (3.3):

$$\Pr(\uparrow, x | \uparrow, z) = \cos^2(\pi/4) = \Pr(\downarrow, x | \uparrow, z) = \sin^2(\pi/4) = 1/2. \quad (3.4)$$

Now consider an atom from the $\uparrow$ -output channel of the *second* apparatus. How should we describe it? In analogy with the first apparatus, we might offer a description of the form $|\uparrow, x\rangle$, i.e., such an atom is fully aligned along the positive x -direction. This would mean that the information that was present at the second apparatus's *input*, where these atoms were aligned in the positive z -direction, was somehow erased. Is that the right approach?

Note that the situation at the input of the first apparatus is different from that at the input of the second apparatus. For the first apparatus, the input consists of the atoms from the oven. It makes sense to consider these to be randomly scrambled in all sorts of ways, including any possible magnetic dipole orientation; maybe the first apparatus somehow merely sorts these orientations, amplifying minute differences. But the input to the second apparatus consists of perfectly z -aligned atoms. We can easily check this by feeding them into a second z -aligned apparatus, as described above. Is any of that z -alignment preserved at the *output* of the second apparatus, given that second apparatus is oriented in the x -direction?

The way to check for this is to add a *third* Stern-Gerlach apparatus, oriented along the z -direction, that takes as input the upper output channel of the second, x -aligned device (Fig. 3.2c). The question is whether any of the alignment along z effected by the first apparatus shows up in what happens at the third apparatus. There are two options. If we find an imbalance of the intensities of the output channels of the third apparatus, we may interpret this as a residue of information from the z -orientation effected by the first apparatus. If, however, the two output channels have equal intensity, then we were right in describing the output in the upper channel of the second experiment as simply $|\uparrow, x\rangle$, with no trace of any previous z -alignment.

And this latter result is, in fact, what we find: the output beams of the third, z -aligned apparatus have equal intensity.²¹ For an atom that made it through the upper channel of the first apparatus (associated with spin state $|\uparrow, z\rangle$) and then the upper channel of the second (associated with spin state $|\uparrow, x\rangle$), the probability of arriving at the upper or lower channel of the third, z -aligned apparatus is exactly the same, $1/2$. The second apparatus has thus wiped out the atom's history of $|\uparrow, z\rangle$, and we were indeed justified in replacing its description by $|\uparrow, x\rangle$ after the second device. Consequently, we will describe the state of an atom from the upper channel of the third experiment simply as $|\uparrow, z\rangle$. In Stern-Gerlach experiments consisting of multiple devices in sequence, the outcome of the *last* interaction is all that counts when it comes to spin alignment.

The Stern-Gerlach Experiment as Measurement of Spin We will provide a more detailed, foundations-focused discussion of the Stern-Gerlach experiment in Sect. 3.3.3, where we will describe the experiment as the measurement of spin as an intrinsic quantum property of the silver atoms. The experiment provides an ideal background for such a discussion because the dynamics of the interaction of the atoms with the inhomogeneous magnetic field have been studied in much detail. That discussion will underline the importance of irreversible interactions in quantum measurements.

But first, we discuss the mathematical framework of QM that was worked out during a brief time in the mid-1920s and which has been astonishingly stable since then.

3.2 The Mathematical Framework of QM

In stark contrast to the continuing debate about the conceptual foundations of quantum mechanics, some of which we will address further below, the theory's mathematical framework appears to stand very firm. Briefly, the main lines are as follows.²²

²¹ To the best of our knowledge, the experiments described here—using silver atoms and two or three actual Stern-Gerlach devices—have not been performed. Historically, the method of adding a static or oscillatory magnetic field, as invented by Rabi et al. (1938), was used instead, and more recent experiments use different systems that are much easier to handle than atomic beams of silver atoms. The conceptual message stays the same.

²² For good introductions to QM and its mathematical structure, see, e.g., the textbooks mentioned in Footnote 10. See also Ismael (2025), which includes a helpful guide to current quantum mechanics textbooks that is aimed at a philosophical audience.

3.2.1 Mathematical Basics: Hilbert Space

Given a physical system of a specific type (e.g., an elementary particle or an atom, or a larger composite system), QM posits that the possible states of that system can be represented in some appropriate *Hilbert space*, a complex vector space with the standard inner product $(\cdot, \cdot)$ defined on it. Depending on the system and the degrees of freedom one intends to model, the dimensionality of that space can be finite or infinite. For example, a realistic model of a silver atom in a Stern-Gerlach apparatus requires an infinite-dimensional Hilbert space (because the position degree of freedom is continuous), but to model the spin degree of freedom that is relevant for an atom's alignment in a magnetic field, one only requires a Hilbert space of dimension $d = 2$. Many specifically quantum phenomena are already exhibited by systems with $d = 2$ (so-called two-state systems, or qubits); some further peculiarities of QM show up in systems of $d = 3$ or higher. We will hardly ever have to consider infinite-dimensional Hilbert spaces in this book, and we will not introduce the full mathematical formalism for working with them.

The bra-ket Formalism In our description of the Stern-Gerlach experiment we already employed, without much comment, an important part of a versatile and useful formalism for QM, namely, Dirac's bra-ket formalism. Let us fix some Hilbert space $\mathcal{H}$, e.g., the two-dimensional complex vector space that is adequate for the relevant spin degree of freedom of a silver atom. The *ket* $|\uparrow, z\rangle$, which we used to describe the state of an atom coming out of the upper channel of a z -oriented Stern-Gerlach apparatus, represents a (normalised) vector in that space. Dually, a *bra*, e.g., $\langle\uparrow, x|$, represents a vector of the dual space (the space of linear functionals on the given Hilbert space, which map vectors to complex numbers). A *bra-ket*, such as $\langle\uparrow, x|\uparrow, z\rangle$, then conveniently represents the inner product of the two vectors $|\uparrow, x\rangle$ and $|\uparrow, z\rangle$ from the Hilbert space, $(|\uparrow, x\rangle, |\uparrow, z\rangle)$, or the application of the linear functional $\langle\uparrow, x|$ on the vector $|\uparrow, z\rangle$. The result is a complex number; for our example, we have $\langle\uparrow, x|\uparrow, z\rangle = 1/\sqrt{2}$, as we will see.

Importantly, the duality between ket and bra space involves complex conjugation, which we write using a superscript asterisk (thus, for $c = a + bi$ with $a, b \in \mathbb{R}$, we have $c^* = a - bi$). If $\langle\varphi|$ is the bra corresponding to ket $|\varphi\rangle$ and c is a complex number, then the bra corresponding to the ket $c|\varphi\rangle$ is $c^*\langle\varphi|$. For the inner (or scalar) product, we have the corresponding relation,

$$\langle\chi|\varphi\rangle = \langle\varphi|\chi\rangle^*. \quad (3.5)$$

This implies directly that the inner product of a vector with itself has to be a real number. Furthermore, that number is zero only for $|\varphi\rangle = 0$, the null vector, and is otherwise positive:

$$\langle\varphi|\varphi\rangle \geq 0. \quad (3.6)$$

One can therefore define a *norm* on $\mathcal{H}$ that is positive definite, via

$$\| |\varphi\rangle \| = \sqrt{\langle \varphi | \varphi \rangle}. \quad (3.7)$$

The normalisation condition for a vector can then be written as $\| |\varphi\rangle \| = 1$, or as $\langle \varphi | \varphi \rangle = 1$.

Dirac's formalism also allows one to write *ket-bras*, e.g., $|\varphi\rangle\langle\varphi|$, which usefully represent specific *linear operators* on the Hilbert space. The result of applying such an operator on a ket vector can be evaluated using the built-in associativity of the formalism, e.g.,

$$(|\varphi\rangle\langle\varphi|)|\chi\rangle = |\varphi\rangle(\langle\varphi|\chi\rangle) = c|\varphi\rangle, \quad (3.8)$$

where the complex number c is defined as $c = \langle\varphi|\chi\rangle$. Thus, one sees that the application of the operator $|\varphi\rangle\langle\varphi|$ on a ket vector $|\chi\rangle$ yields a vector parallel to $|\varphi\rangle$ with a modulus and phase corresponding to the inner product of the vectors $|\varphi\rangle$ and $|\chi\rangle$. For a normalised vector $|\varphi\rangle$, therefore, $|\varphi\rangle\langle\varphi|$ is the projection operator onto the subspace spanned by $|\varphi\rangle$, which we can also write as $\hat{P}_{|\varphi\rangle}$. (Here and in the remainder of this chapter, we adopt the convention to put hats over capital letters to indicate operators.) It is easy to check for idempotence, which is a defining property of projection operators:

$$\hat{P}_{|\varphi\rangle}^2 = (|\varphi\rangle\langle\varphi|)(|\varphi\rangle\langle\varphi|) = |\varphi\rangle(\langle\varphi|\varphi\rangle)\langle\varphi| = |\varphi\rangle 1 \langle\varphi| = \hat{P}_{|\varphi\rangle}. \quad (3.9)$$

Basis Vectors A Hilbert space $\mathcal{H}$ of finite dimension N has a *basis* consisting of N orthonormal vectors,

$$|b_1\rangle, \dots, |b_N\rangle. \quad (3.10)$$

Orthonormality means that the basis vectors are normalized ($\langle b_i | b_i \rangle = 1$) and pairwise orthogonal. Using Kronecker's delta symbol ($\delta_{ij} = 1$ if $i = j$, and $\delta_{ij} = 0$ otherwise), this condition can be written as:

$$\langle b_i | b_j \rangle = \delta_{ij}, \quad i, j = 1, \dots, N. \quad (3.11)$$

Any vector $|\varphi\rangle$ from $\mathcal{H}$ can be expressed as a linear combination—or, as one also says, a *superposition*—of arbitrarily chosen basis vectors. This freedom of choice of basis is the heart of the superposition principle. That principle is extremely useful because a suitable choice of basis can simplify many problems. A given ket vector $|\varphi\rangle$, written as a superposition of basis vectors $|b_i\rangle$, looks like this:

$$|\varphi\rangle = c_1|b_1\rangle + \dots + c_N|b_N\rangle, \quad (3.12)$$

where the coefficients are

$$c_i = \langle b_i | \varphi \rangle, \quad i = 1, \dots, N. \quad (3.13)$$

So we have the identity

$$|\varphi\rangle = \sum_{i=1}^N |b_i\rangle \langle b_i | \varphi \rangle. \quad (3.14)$$

The power of the Dirac notation shows in the fact that using associativity and noting that the vector $|\varphi\rangle$ was arbitrary, this expression can also be read as telling us that the *operator*

$$\sum_{i=1}^N |b_i\rangle \langle b_i| = \hat{1}, \quad (3.15)$$

i.e., the sum of the projection operators onto the N orthogonal subspaces of the Hilbert space spanned by the basis vectors is the *identity operator* (the “ $\hat{1}$ ” on the right hand side above, with the hat indicating that it is an operator), as it returns an arbitrary vector $|\varphi\rangle$ unchanged:

$$\left(\sum_{i=1}^N |b_i\rangle \langle b_i| \right) |\varphi\rangle = |\varphi\rangle. \quad (3.16)$$

For normalised $|\varphi\rangle$, the modulus squares of the coefficients $c_i = \langle b_i | \varphi \rangle$ sum to unity:

$$\sum_{i=1}^N |c_i|^2 = \sum_{i=1}^N |\langle b_i | \varphi \rangle|^2 = \langle \varphi | \left(\sum_{i=1}^N |b_i\rangle \langle b_i| \right) | \varphi \rangle = \langle \varphi | \hat{1} | \varphi \rangle = \langle \varphi | \varphi \rangle = 1. \quad (3.17)$$

This fact will enable a reading in terms of probabilities in the context of the Born rule, to be discussed in Sect. 3.3.2.2 below.

Operators, Observables, and Eigenvectors In the mathematical formalism of QM, linear operators play an especially important role. We have already encountered *projection operators* onto subspaces of a Hilbert space, such as $\hat{P}_{|\varphi\rangle} = |\varphi\rangle \langle \varphi|$, above. Quite generally, a linear operator $\hat{A}$ maps vectors from a Hilbert space $\mathcal{H}$ to that Hilbert space and satisfies, for arbitrary vectors $|\varphi\rangle$ and $|\chi\rangle$ and for any complex numbers α and β , the linearity condition

$$\hat{A}(\alpha|\varphi\rangle + \beta|\chi\rangle) = \alpha\hat{A}|\varphi\rangle + \beta\hat{A}|\chi\rangle. \quad (3.18)$$

An operator $\hat{A}$ has a *Hermitian adjoint*, written $\hat{A}^\dagger$, defined such that for any ket $|\varphi\rangle$, the ket $\hat{A}|\varphi\rangle$ is in dual correspondence with the bra $\langle \varphi | \hat{A}^\dagger$. An operator is *Hermitian* if

$$\hat{A} = \hat{A}^\dagger. \quad (3.19)$$

In quantum mechanics, any physical observable, that is, any measurable quantity of a physical system, is mathematically described by a Hermitian operator (also called *observable*), on its associated Hilbert space. In the case of a finite-dimensional Hilbert space that is at issue here, the reverse is often also true, i.e., any Hermitian operator corresponds to some physical observable.²³

Generally, a ket $|\varphi\rangle$ (i.e., a vector in our N -dimensional Hilbert space $\mathcal{H}$) is called an *eigenvector* of operator $\hat{A}$ if

$$\hat{A}|\varphi\rangle = c|\varphi\rangle \quad \text{for some } c \in \mathbb{C}; \quad (3.20)$$

in this case, the complex number c is called the corresponding *eigenvalue*. One often highlights the eigenvector-eigenvalue correspondence by naming an eigenvector via the eigenvalue, e.g., $\hat{A}|a\rangle = a|a\rangle$. One can show quite easily that given a Hermitian operator $\hat{A}$, its eigenvalues are real, and the Hilbert space $\mathcal{H}$ has an orthonormal basis consisting of eigenvectors of that operator.²⁴

$$\begin{aligned} |a_1\rangle, \dots, |a_N\rangle; \quad \hat{A}|a_i\rangle &= a_i|a_i\rangle; \quad a_i \in \mathbb{R}; \\ \langle a_i|a_j\rangle &= \delta_{ij}; \quad i, j = 1, \dots, N. \end{aligned} \quad (3.21)$$

In that basis, we can represent the observable as a matrix in diagonal form, and the entries on the diagonal are then the respective eigenvalues. This corresponds to the spectral decomposition of the observable

$$\hat{A} = \sum_{i=1}^N a_i |a_i\rangle \langle a_i|. \quad (3.22)$$

It need not be the case that all the a_i are different; i.e., an eigenspace (the subspace of Hilbert space spanned by eigenvectors corresponding to a given eigenvalue) can have a dimension larger than one. (For a trivial example, consider the identity operator, $\hat{1}$: any vector is an eigenvector of that operator, with eigenvalue 1; so the corresponding eigenspace has dimension N .) But eigenvectors corresponding to different eigenvalues are necessarily orthogonal, i.e., the different eigenspaces are disjoint.

A weighted sum of Hermitian operators is also Hermitian. In particular, given a set of M weights (positive real numbers) that sum to one, $w_1, \dots, w_M$ s.t.

²³ The infinite-dimensional case is subtle. Often a Hermitian operator corresponds to an observable if there is a basis of the Hilbert space consisting solely of its eigenvectors, i.e., if the set of eigenvectors is complete. (While in a finite-dimensional Hilbert space every Hermitian operator has a complete set of eigenstates, this is not true in the infinite dimensional case.) For further details see, e.g., Cohen-Tannoudji et al. (2019) and Ballentine (2015).

²⁴ For a proof, see, e.g., Sakurai and Napolitano (2020, Ch. 1.3). Things are a bit more complicated for infinite-dimensional Hilbert spaces; see Footnote 23 above.

$\sum_{i=1}^M w_i = 1$, and M unit vectors $|\varphi_1\rangle, \dots, |\varphi_M\rangle$, one can define the Hermitian operator

$$\rho = \sum_{i=1}^M w_i |\varphi_i\rangle \langle \varphi_i|. \quad (3.23)$$

The *trace* of an operator is defined as

$$\text{tr}(\hat{A}) \equiv \sum_{i=1}^N \langle b_i | \hat{A} | b_i \rangle, \quad (3.24)$$

where $\{|b_i\rangle\}_{i=1,\dots,N}$ is an arbitrary orthonormal basis. Using Eq. (3.17) and the fact that the weights w_i sum up to unity, it is easy to show that the trace of an operator of the form (3.23) is one, $\text{tr}(\rho) = 1$. Note that we did not impose the condition that $M = N$ (there can be fewer or more terms in the sum than the dimension of the Hilbert space), nor that $\langle \varphi_i | \varphi_j \rangle = \delta_{ij}$ (the vectors need not be pairwise orthogonal; they just need to be normalised). An operator ρ of the form (3.23) is called a *density operator*. Density operators play an important role in the description of quantum systems under less-than-perfect information and of parts of composite quantum systems. For example, the atoms coming out of the oven in the Stern-Gerlach experiment are adequately described by a density operator that carries no information about any direction of spin alignment (see our discussion in Sect. 3.3.3).

Apart from projection operators, observables, and density operators, there is a further class of linear operators that are highly important in the mathematical framework of QM: the *unitary operators*. An operator $\hat{U}$ is *unitary* if it is invertible and its inverse is its Hermitian adjoint, i.e.,

$$\hat{U} \hat{U}^\dagger = \hat{1} \quad \text{and} \quad \hat{U}^\dagger \hat{U} = \hat{1}. \quad (3.25)$$

Unitary operators are, thus, automorphisms of the Hilbert space, mapping the whole space onto itself in a structure-preserving way. They therefore also preserve the inner product between vectors: for

$$(|\varphi\rangle, |\chi\rangle) = \langle \varphi | \chi \rangle = c, \quad (3.26)$$

we have

$$(\hat{U}|\varphi\rangle, \hat{U}|\chi\rangle) = \langle \varphi | \hat{U}^\dagger \hat{U} | \chi \rangle = \langle \varphi | \hat{1} | \chi \rangle = \langle \varphi | \chi \rangle = c. \quad (3.27)$$

Rotations are typical examples of such automorphisms. Note that, apart from the trivial case of the unit operator, projection operators—such as $P_{|\uparrow, z\rangle}$ —are not unitary: they map the whole Hilbert space onto a proper subspace and, thus being non-injective, cannot be inverted.

Implicit in the above relations is the fact that the linear operators on a given Hilbert space can be not just added, but also multiplied, and that there is a neutral

element of multiplication, the unit operator $\hat{1}$. The operator algebra is, for the finite-dimensional case at issue here, isomorphic to the familiar algebra of $N \times N$ matrices over the complex numbers. (For this reason, Heisenberg's first consistent formulation of QM from 1925 was known as "matrix mechanics".) It is a general fact of matrix multiplication, which also applies to linear operators, that while multiplication is associative:

$$(\hat{A}\hat{B})\hat{C} = \hat{A}(\hat{B}\hat{C}), \quad (3.28)$$

multiplication need not be commutative. That is, the product of two operators $\hat{A}$ and $\hat{B}$ in one order ($\hat{A}\hat{B}$) need not equal the product of these operators in the reverse order ($\hat{B}\hat{A}$). One defines the *commutator* of two operators, written via square brackets, as the residual difference:

$$[\hat{A}, \hat{B}] \equiv \hat{A}\hat{B} - \hat{B}\hat{A}. \quad (3.29)$$

If the commutator $[\hat{A}, \hat{B}] = 0$, we say that the operators *commute*. It can be shown that if two observables (Hermitian operators) commute, there is a basis of the Hilbert space that consists of joint eigenvectors of both $\hat{A}$ and $\hat{B}$. In case $[\hat{A}, \hat{B}] \neq 0$, on the other hand, one says that the operators A and B *do not commute*. Non-commuting observables (i.e., non-commuting Hermitian operators, for which there is no basis of joint eigenvectors) are at the heart of genuinely quantum phenomena such as complementarity and the uncertainty relation, to be discussed below.

This is, of course, only the beginning of a description of the mathematics of QM. We will not work out further properties or relations based on the material just given; these are covered in many excellent textbooks.²⁵ Rather, we will now discuss the use of the mathematical framework in the application of QM to real physical systems.

3.2.2 Interpreting the Formalism: The Basics

The mathematical formalism of QM was developed with a clear aim in mind: it was meant to be useful for describing, both qualitatively and quantitatively, what to expect when one studies and manipulates quantum systems. The theory of QM is, first and foremost, a tool that human agents—starting with the physicists of the 1920s—constructed to find their way among the many challenging and confusing phenomena of the quantum world. That tool had to be developed in tandem with an interpretation that attaches meanings to its different mathematical parts that help one to apply the tool skilfully.

²⁵ Apart from the references listed in Footnotes 10 and 22, textbooks with a focus on the mathematics of QM include, historically, Von Neumann (1932) and, more recently, Galindo and Pascual (1990) and Gustafson and Sigal (2011).

So much, we hope, is uncontroversial—in fact, this holds for any physical theory. But the actual interpretation of the QM formalism is the subject of many controversies. It is probably impossible not to trigger such controversies when one writes about fundamental issues of QM, but for now, we try to describe just the basics and to keep our commitments minimal.

The Hilbert Space and Its Elements We said above that a Hilbert space $\mathcal{H}$ of dimension N is chosen depending on which system one wants to consider, and which degrees of freedom of that system are deemed relevant for modelling a situation of interest. Relative to that selection of degrees of freedom, there are states of maximal information about the system. These maximally specific, so-called *pure states* are adequate for describing systems in cases in which they are known to be suitably precisely prepared. For example, the atoms coming out of one of the channels of an ideal Stern-Gerlach device are, with respect to their spin degree of freedom, in a pure state.

Pure states are represented by one-dimensional subspaces or, equivalently, by normalised vectors in the Hilbert space $\mathcal{H}$, and are written as kets, e.g., $|\uparrow, z\rangle$. There is also an associated formalism for handling non-pure state ascriptions under less-than-perfect information, via so-called density operators ρ , touched upon above, which we will discuss further below. So far, this is quite similar to classical theories, in which a system's state is also represented via some abstract space or other (e.g., by a point in phase space), with associated tools for handling less-than-perfect information (e.g., via phase-space density functions). The crucial difference lies in what one can say about the various possible properties that the system can exhibit. As we will see below, in QM it is not the case that a system's state provides unique values for all these possible properties.

Attributing a Quantum State How do we assign a state to a system; what do we write in a ket, or which density operator do we use? Early in the development of QM it was quite customary to think of the attribution of quantum states as pertaining to ensembles of systems, e.g., the ensemble of the atoms that come out of the oven in the Stern-Gerlach experiment. It has become clear, however, that in many (and in our view, in the most interesting) cases, an ensemble view is not adequate. With current technologies, matter can be manipulated at the level of individual atoms, and QM applies to such systems. Quantum states are, therefore, attributed individually to single systems such as single atoms. One is often in a position to attribute a quantum state—a density operator or even a pure state, a ket—if one has just *measured* a system, or if the system has been *prepared* using a source with known properties.

Before we discuss the general case involving density operators, we focus on pure states for a while, as their mathematics is somewhat simpler.

Attributing Another Quantum State In our example of the two-stage (first z , then x) Stern-Gerlach experiment, the original source—the oven—produces atoms that have to be assigned a maximally impure spin state: nothing about their spin orientation is known. After the measurement, or preparation, of the spin component along the z -direction by the first (z -oriented) Stern-Gerlach device, it is possible to

attribute a pure spin state: the “spin up along z ” state, $|\uparrow, z\rangle$, for those in the upper channel, and the “spin down along z ” state, $|\downarrow, z\rangle$, for those in the lower channel. One can consider the effect of the first apparatus as a measurement of the atoms’ spin projection along the z axis, or one can consider the upper channel of that first apparatus as a source of atoms prepared in the pure state $|\uparrow, z\rangle$ (and similarly for the lower channel). Either way, one is in a position to attribute a pure state to a single atom, based on the information that one has about that atom (in this case, which output channel it is from). Now, when that atom is sent through the second, x -oriented apparatus, this amounts to a measurement of spin projection along the x -axis, and the outcome is, as we said, purely random, like a coin flip. To those atoms that come out of the upper channel of the second apparatus, we can then assign the updated pure state $|\uparrow, x\rangle$, and to those from the lower channel, $|\downarrow, x\rangle$.

There are two questions one can ask about these updates of state assignments. First, how well can one predict which update one will have to make? And second, what happens, physically, at such an update, and what happens in between consecutive updates?

These questions go to the heart of the peculiarities of QM. As to the first question: Classically, one should expect that if one has a maximal amount of information about a system (if one has assigned a pure state), one should be able to make predictions with certainty—unless the dynamics is indeterministic, but this is classically quite rare.²⁶ In QM, however, our above description suggests that even in the case of an atom in a pure state such as $|\uparrow, z\rangle$, one cannot do better than chance if one wants to predict which channel the atom comes out after interaction with a subsequent, x -oriented Stern-Gerlach apparatus (or, as one says, what will be the result of measuring its spin projection along the x -axis). Sure enough, one can predict with certainty what happens if the second apparatus is oriented in the z -direction—the pure state $|\uparrow, z\rangle$ is confirmed—but one has no predictively useful information at all about the case of a second device that is x -oriented. How and why is that?

There is a clear urge, felt by many of the people working on QM in the early days (most notably, by Einstein), to think that the state assignment must be incomplete: $|\uparrow, z\rangle$ would thus not be a fully specified pure state after all, but just a state of partial information that could, at least in principle, be fine-grained by additional information, values of so-called hidden variables. We will discuss the dialectics of this urge further below, under the heading of “the deterministic itch” (Sect. 3.4). If one rejects this urge, as one should, then one has to acknowledge that in most cases, the best one can do with respect to the systems studied by QM is to make *probabilistic* predictions. One has to wait and see and do the update in accordance with what one finds, and that is when a new state can be assigned; prior to that, one

²⁶ While not the norm, there *are* classical systems whose dynamics is indeterministic. For a simple case, consider a collision of three point masses of Newtonian mechanics. Further, more involved examples are described in detail in Earman (1986). A recently much discussed example is what has become known as Norton’s dome, see Norton (2008).

can only assign probabilities to the different possible updates that one may be forced to make. This, of course, makes our second question all the more salient: What happens to the system at the time when one updates its state, and what happens in between the attributions of different quantum states at different times? More generally, what happens to a quantum system as time progresses? This is the subject of quantum dynamics, including the dynamics of quantum measurements (an aspect of the so-called “measurement problem”). We will now describe two different ways of approaching this topic within the undisputed, canonical mathematical framework of QM, the so-called Schrödinger and Heisenberg pictures of quantum dynamics. In the end, we will argue that the Heisenberg picture should be preferred, because that picture can be used to avoid conceptual difficulties by showing how QM is a theory for and by agents.

3.2.3 Quantum Dynamics: Schrödinger vs. Heisenberg Picture

As we just said, there are two different approaches to quantum dynamics, both regularly taught in standard courses. While these approaches are equivalent when it comes to computing probabilistic predictions and expectation values for observable effects, they are conceptually quite different. The approaches are named after two men who worked out competing, but ultimately equivalent ways of describing QM in the mid-1920s, Schrödinger and Heisenberg. Their approaches do not only paint a different picture of what happens to a quantum system’s state “when nobody is looking”, but they also come with a different take on what happens *when* somebody is looking, e.g., via measurement.

The Schrödinger Picture In the Schrödinger picture, one views the instantaneous quantum state of a system (for simplicity, here a pure state like $|\psi(t_0)\rangle$ at time t_0) as evolving over time subject to a dynamical equation—the famous *Schrödinger equation*—that is modelled after familiar classical dynamics. In that equation, $\hat{H}$ is the Hamilton operator, a Hermitian operator—an observable—that describes the system’s total energy:

$$i\hbar \frac{\partial}{\partial t} |\psi(t)\rangle = \hat{H} |\psi(t)\rangle. \quad (3.30)$$

This equation is linear, which makes the superposition principle of Eq. (3.12) especially pertinent: under these dynamics the time evolution of a superposition (linear combination) of basis states is equal to the superposition of the individually evolved basis states. If these states are properly chosen, solving Eq. (3.30) can be greatly simplified. For example, when $\hat{H}$ does not depend on time, it is useful to choose a basis consisting of eigenstates of $\hat{H}$, i.e., states whose energy remains constant over time.

Integrating the Schrödinger equation, one can derive a unitary time evolution operator $\hat{U}(t_1, t_0)$ that relates the system state at an arbitrary time t_1 to the state at a given time t_0 :

$$|\psi(t_1)\rangle = \hat{U}(t_1, t_0)|\psi(t_0)\rangle. \quad (3.31)$$

In the case that $\hat{H}$ does not depend on time, just mentioned above, one has the simple expression²⁷

$$\hat{U}(t_1, t_0) = e^{-\frac{i}{\hbar}\hat{H}(t_1-t_0)}. \quad (3.32)$$

This dynamics implies that even at times at which no information about the system is collected, the system state is continuously updated, traversing a trajectory in the abstract Hilbert space of possible pure states. The dynamics is called the “unitary dynamics” because $\hat{U}$ is unitary. In general, it cannot describe what happens at a state update, as such an update is only probabilistically determined. Accordingly, a second type of (non-unitary) dynamics is invoked, the (in-)famous “collapse of the wave function” upon measurement. Roughly, updating the system’s state after measurement (say, at time t_1) amounts to a reset of the dynamics, by instantaneously replacing the state that was evolved according to the above dynamics up to the time of the measurement, $|\psi(t_1)\rangle = \hat{U}(t_1, t_0)|\psi(t_0)\rangle$, with a new state assignment that accounts for the actually measured value.²⁸ The so-called measurement problem of QM consists of two hard questions about this account: how to live with this dichotomy of two incompatible dynamics, and how to describe the physical process of measurement in a uniform and consistent way, given that it amounts to a dramatic disruption of the well-behaved dynamics according to Eq. (3.30). None of these questions has so far been answered in a fully satisfactory way.²⁹

²⁷ The exponentiation of the operator $\hat{H}$ can be defined via its spectral decomposition, $\hat{H} = \sum_{i=1}^N E_i |E_i\rangle\langle E_i|$, where $|E_i\rangle$ are the eigenvectors of $\hat{H}$ and E_i the corresponding eigenvalues, i.e., the possible energies of the system. One then straightforwardly obtains $\hat{U}(t_1, t_0) = \sum_{i=1}^N e^{-\frac{i}{\hbar}E_i(t_1-t_0)} |E_i\rangle\langle E_i|$.

²⁸ As we will discuss in Sect. 3.3.2.2 below, that new state is the projection of $|\psi(t_1)\rangle$ onto the eigenspace of the measured observable corresponding to the measured value. In the simplest case, this amounts to replacing $|\psi(t_1)\rangle$ by one of the eigenstates corresponding to the possible measurement outcomes.

²⁹ With a view to resolving the dynamics of the measurement process, one can include the interaction of the measurement apparatus with the system into the description. As an example, this could be the interaction of a laser field with the internal states of an atom, as used in conditional resonance fluorescence techniques for atomic state detection in quantum optics experiments. Using such a description, one can show how the density matrix describing the atomic system alone continuously evolves from an initially pure state (projector) into a diagonal density matrix $\rho = \sum_{i=1}^N p_i |E_i\rangle\langle E_i|$, that is, a mixed state with $|E_i\rangle$ denoting the eigenstates of energies E_i as in Footnote 27 and p_i the probabilities with which the different energies E_i are found in the measurement. Such an approach cannot, however, further resolve how one of the different energies will be selected.

The Heisenberg Picture In the Heisenberg picture, in contrast to a continuous update of the system's state $|\psi(t)\rangle$, the assignment of a quantum state $|\psi\rangle$ is always dependent on the availability of information about the system. This could be the characteristics of a particle source such as the oven in the Stern-Gerlach apparatus (in this case, the information does not suffice to assign an individual atom a pure spin state), or knowledge about which channel of a Stern-Gerlach device an atom ended up in (in this case, a pure spin state such as $|\uparrow, z\rangle$) can be assigned). Any state assignment will thus have to be based on the information that is actually available, and an updated state is assigned not continuously, but only at specific times determined by the chosen model of the system and its interactions—for example, after the registration of a measurement outcome, or based on knowledge of some other irreversible physical process with the right characteristics.

It is sometimes useful to put a time subscript on an assigned state, to keep track of when the state was assigned, as in $|\psi\rangle_{t_0}$, but the t_0 in that notation is not a parameter along which the state changes, but just a mnemonic helper. In between the times at which information becomes available, the state assigned to the system is kept constant.³⁰ Not evolving over time, the quantum state in the Heisenberg picture has a conceptually different status than the time-dependent state in the Schrödinger picture: It is even less plausible to view the quantum state as something that is simply out there in nature—rather, in the Heisenberg picture, the state belongs to the one who assigns it, an agent who uses QM.

Of course, there is change, which needs to be accounted for, no matter which picture is chosen. But in the Heisenberg picture, the dynamics of what happens in nature is described not via state evolution, but rather, as a change over time in the observables that represent the physical quantities that a user of the theory is interested in. Here, there are strong similarities to classical dynamics. The dynamics of these observables is described by the *Heisenberg equation of motion*,

$$\frac{d}{dt}\hat{A}(t) = \frac{1}{i\hbar}[\hat{A}(t), \hat{H}(t)] + \frac{\partial \hat{A}(t)}{\partial t}, \quad (3.33)$$

which states that the rate of change of an observable $\hat{A}$ is given by one component that is proportional to the commutator of that observable and the system's Hamiltonian, and a second component that, if it does not vanish, describes the explicit (or parametric) time dependence of the observable itself. With the dynamics thus attributed exclusively to the observables, the quantum state ascription in the Heisenberg picture is only changed through explicit updates, and consequently, state ascription is always with respect to an agent who uses the theory.

Note that, as in classical mechanics, in quantum mechanics any observable of a given system—for example, of a free particle or a spin-1/2 system—can be written as a function of its dynamical variables, which are certain independent basic

³⁰ Also, instead of $|\psi\rangle$ one often writes specific information into the ket, e.g., regarding the observable that was measured and the result that was obtained in the measurement, as in Eq. (3.46).

observables of the system—in our examples, position and momentum, or a pair of complementary spin components, respectively. Similarly, the density operator $\rho(t)$ defined in Eq. (3.23) can be written as a function of the system’s dynamical variables³¹ and, accordingly, it satisfies Eq. (3.33), which then takes the form

$$\frac{d}{dt}\rho(t) = \frac{1}{i\hbar}[\rho(t), \hat{H}(t)] + \frac{\partial\rho(t)}{\partial t} = 0. \quad (3.34)$$

The last equality states that the information that an agent assigns to a system does not change over time. It can be written in the form

$$\frac{\partial\rho(t)}{\partial t} = \frac{1}{i\hbar}[\hat{H}(t), \rho(t)], \quad (3.35)$$

which is also referred to as the von Neumann equation for the density operator. We will elucidate the meaning of the partial derivative below, in the context of Eq. (3.41).

Heisenberg’s equation of motion for the density operator corresponds to the Liouville equation in classical statistical mechanics, with the commutator replaced by the so-called Poisson bracket (Goldstein et al., 2002). The constancy of the density operator that contains the agent’s information about the system (e.g., due to its preparation at a certain time t_0), $\rho(t) = \rho(t_0)$, similarly translates into Liouville’s theorem in classical statistical mechanics: The explicit time dependence of the distribution function in phase space only compensates for the implicit time dependence in the canonical dynamical variables.³²

Empirical Equivalence vs. Conceptual Difference of the Two Pictures In comparing the two pictures, we will first show their empirical equivalence for matters of computing empirically accessible quantities like expectation values, before we consider their conceptual differences more closely. For simplicity, we assume in the following that the system Hamiltonian $\hat{H}$ is constant over time, and that the observable $\hat{A}$ that we are interested in has no explicit time dependence. These assumptions hold, e.g., for problems of determining atomic spectra, for binding problems in physical chemistry, or for many problems in solid state physics.

³¹ For a clear exposition of this approach, see, e.g., Englert (2006b, Ch. 1.3).

³² In summary, in classical statistical mechanics, the phase-space density thus plays a role analogous to the density operator in quantum mechanics in that it can summarise our statistical description of an ensemble of systems, but it can also summarise our knowledge of a single system. Even though the ontology assumed in classical mechanics may be different from quantum mechanics, the separation of an epistemic part (phase space density) from the dynamical part (dynamical variables) is similar in classical mechanics and in quantum mechanics in the Heisenberg picture (Auletta et al., 2009). This analogy with classical statistical mechanics, which fits well with the agency perspective we are developing in this book, speaks in favour of the Heisenberg picture as a natural choice.

This means that we can integrate the Schrödinger equation (3.30) to yield an explicit, simple expression for the unitary time evolution operator, see Eq. (3.32). Starting with a state assignment $|\psi\rangle_0$ at time $t_0 = 0$ (and dropping the second argument from $\hat{U}$ for convenience), we can calculate the expectation value for observable $\hat{A}$ at time t in the Schrödinger picture (superscript S) as

$$\langle \hat{A}^S \rangle_t = \langle \psi(t) | \hat{A}^S | \psi(t) \rangle = \langle \psi |_0 \hat{U}^\dagger(t) \hat{A}^S \hat{U}(t) | \psi \rangle_0. \quad (3.36)$$

In the Heisenberg picture, in order to evaluate the expectation value, we use the fixed quantum state assigned at $t_0 = 0$, $|\psi\rangle_0$, and compute the Heisenberg picture observable $\hat{A}^H(t)$ for the time we are interested in. Under our assumptions, the Heisenberg equation of motion can be integrated, starting from $\hat{A}^H(0) = \hat{A}^S$, to yield

$$\hat{A}^H(t) = \hat{U}^\dagger(t) \hat{A}^S \hat{U}(t), \quad (3.37)$$

which gives us

$$\begin{aligned} \langle \hat{A}^H \rangle_t &= \langle \psi |_0 \hat{A}^H(t) | \psi \rangle_0 = \langle \psi |_0 \hat{U}^\dagger(t) \hat{A}^S \hat{U}(t) | \psi \rangle_0 \\ &= \langle \psi(t) | \hat{A}^S | \psi(t) \rangle = \langle \hat{A}^S \rangle_t. \end{aligned} \quad (3.38)$$

So much for expectation values, what about the eigenvalues of $\hat{A}^S$ vs. $\hat{A}^H(t)$, which correspond to the possible observations (measurement outcomes)? It is easy to check that if $|a_i\rangle$ is an eigenvector of $\hat{A}^S$, so that $\hat{A}^S |a_i\rangle = a_i |a_i\rangle$ for the eigenvalue a_i , then the transformed ket $|a_i, t\rangle = \hat{U}^\dagger(t) |a_i\rangle$ is an eigenvector of $\hat{A}^H(t)$ with the same eigenvalue, and vice versa:

$$\begin{aligned} \hat{A}^H(t) |a_i, t\rangle &= \hat{U}^\dagger(t) \hat{A}^S \hat{U}(t) |a_i, t\rangle = \hat{U}^\dagger(t) \hat{A}^S \hat{U}(t) \hat{U}^\dagger(t) |a_i\rangle \\ &= \hat{U}^\dagger(t) \hat{A}^S |a_i\rangle = \hat{U}^\dagger(t) a_i |a_i\rangle = a_i |a_i, t\rangle. \end{aligned} \quad (3.39)$$

Lastly and relatedly, the coefficients c_i for expanding the quantum state in terms of eigenvectors of an observable $\hat{A}$, which ground the probabilistic predictions of quantum mechanics according to the Born rule (see Sect. 3.3.2.2 below), are also unchanged: For $\{|a_i\rangle\}_{i=1,\dots,N}$ a basis of our N -dimensional Hilbert space consisting of eigenvectors of the observable $\hat{A}^S = \hat{A}^H(0)$ in question, in the Schrödinger picture we can write

$$|\psi(t)\rangle = \sum_{i=1}^N c_i^S(t) |a_i\rangle, \quad \text{where} \quad c_i^S(t) = \langle a_i | \psi(t) \rangle = \langle a_i | \left(\hat{U}(t) | \psi \rangle_0 \right). \quad (3.40)$$

In the Heisenberg picture (noting that as the eigenkets $|a_i\rangle$ transform as $\hat{U}^\dagger(t)|a_i\rangle$, the eigenbras $\langle a_i|$ transform as $\langle a_i|\hat{U}(t)$), we have

$$|\psi\rangle_0 = \sum_{i=1}^N c_i^H(t) |a_i, t\rangle, \quad \text{where} \quad c_i^H(t) = \langle a_i, t | \psi \rangle_0 = \left(\langle a_i | \hat{U}(t) \right) |\psi\rangle_0. \quad (3.41)$$

The meaning of the partial derivative $\frac{\partial}{\partial t}$ in the von Neumann Eq. (3.35) can now be elucidated by considering the special case of a pure state, $\rho = |\psi\rangle_0 \langle \psi|$, where $|\psi\rangle_0$ is constant in time: The partial derivative $\frac{\partial}{\partial t}$ refers to the time dependence of the expansion coefficients $c_i^H(t)$ in Eq. (3.41) that compensates for the dynamical time dependence inside the basis kets $|a_i, t\rangle$, which in turn is inherited from the dynamical time dependence of the observable $\hat{A}(t)$ that has $|a_i, t\rangle$ as its eigenstates.

To see the empirical equivalence of the pictures, note that from Eq. (3.41), by associativity, we have

$$c_i^H(t) = \left(\langle a_i | \hat{U}(t) \right) |\psi\rangle_0 = \langle a_i | \left(\hat{U}(t) |\psi\rangle_0 \right) = c_i^S(t). \quad (3.42)$$

From this mathematical perspective, then, the switch between the pictures appears to amount just to reading the time evolution operator $\hat{U}(t)$ as contributing to a dynamically evolving quantum state $|\psi(t)\rangle$, or to a dynamically evolving observable $\hat{A}^H(t)$ and, consequently, a dynamically evolving basis of eigenstates. All values that are linked to observable system behaviour—the eigenvalues a_i and the coefficients $c_i(t)$ as well as expectation values—remain the same.

Despite their empirical equivalence, though, these two approaches to quantum dynamics suggest radically different views on the status of the quantum state, and accordingly, on many other foundational issues of QM. We will discuss these in the context of applying QM in Sect. 3.3.

3.3 Applying QM: When to Update the Quantum State

When one uses QM in a concrete application, one has to pick the right model of the system one is dealing with together with a useful interpretation of the model's parts. Mathematically, a model of the system is often given by the specific form of the Hamilton operator, which may contain, for example, the electromagnetic interactions of different parts of the system with each other and with external (control) fields. Another crucial element in this is to take a stance on when the quantum state ascribed to a system should be changed or updated. The two mentioned pictures of quantum dynamics come with fundamentally different commitments in relation to state updates.

3.3.1 *Picking a Picture*

In the Schrödinger picture, as we said, one takes a quantum system to have a state that evolves over time like the state of a classical system. For a classical system, the state is supposed to provide full information about the system in all relevant respects: any property of the system should be expressible as a function of the state.³³ Having a dynamically evolving quantum state invites one to take the quantum state to be an objective feature of the world out there that, at any given time, represents the whole truth about the system. This view, however, is in obvious tension with the fundamentally probabilistic nature of the predictions for state updates that QM makes.

In the Heisenberg picture, a system is assigned a state (a density operator ρ or a state ket $|\psi\rangle$) only when information about the system being in that state is actually available.³⁴ This is normally the case after a suitable physical interaction, which may be the preparation of the system via a source of known properties, or a measurement with a determined outcome. Such a state is, at least implicitly, always ascribed by an agent who uses the theory of QM. Unless that user of the theory has relevant new information, it is not adequate to update the description of the system. As we just saw, however, there is a way of projecting forward the expected dynamics that the system is subject to (by integrating Eq. (3.33)), so that it is possible to make—probabilistic—predictions about the system for later times, e.g., when a system property will be measured. These predictions, which match the predictions of calculations based on the Schrödinger picture, can then be compared with actual new information, and that new information can be used to update the state assignment.³⁵

Both approaches have, of course, their costs and benefits. These do not so much show up in the practicalities of actual calculations, because the mathematics involved is quite similar. The costs and benefits are, rather, of a conceptual nature. It seems fair to say that much of the quantum foundations literature is wedded to the Schrödinger picture, with its classical idea of system states evolving continuously over time. A user of QM in the Schrödinger picture can retreat to the position of a mere observer—unless, that is, the system undergoes measurements, which

³³ The state is thereby taken to be a maximal intrinsic characterisation of the system, so that it should ground all properties and relations. Philosophical issues arise already at this stage, e.g., in relation to the mathematical surplus of state assignments in theories with gauge freedom. See Butterfield (2005) for an overview.

³⁴ In an extreme case in which no information is available, one has to assign a completely mixed state, i.e., a density matrix that is proportional to the identity. In the Stern-Gerlach experiment, such a state describes the spin state of an atom that leaves the oven, before it enters the magnetic field of the apparatus.

³⁵ The original idea of quantum Bayesianism was to view this update as a kind of subjective Bayesian update done by the user of the theory. Current QBism, while putting an agent as the user of QM in the centre as well, has mostly abandoned the Bayesian framework in favour of a broader pragmatist approach by now. See, e.g., Healey (2023) for an overview, and see Chap. 4 for our discussion of QBism.

are disruptive in that picture (as we will describe in more detail below). The Heisenberg picture, on the other hand, is much better suited to an understanding of QM as a theory of agents interacting with nature. The conceptual costs here are those of making sense of agents in a physical world. In the context of this book, we acknowledge and welcome this challenge—it sets the theme for our whole enterprise. But the quantum state is, on that picture, conceptually unproblematic: it is, first and foremost, something that an agent as a user of QM contributes to the overall description, and which she can update depending on the information that becomes available to her.³⁶ Different agents may be in a different position here—and acknowledging that is, in our view, really helpful for understanding the supposed riddles of quantum correlations, which play such a big role in foundational debates (see Sect. 3.4.3 below). Anyway, there is on this picture much less inclination to think of the quantum state as somehow “the whole truth” about a quantum system in question.

In our view, these latter aspects make the Heisenberg picture the more attractive point of view for foundational discussions, especially when agency is to be discussed explicitly, as in this book. We already acknowledged that this is not the majority position. It is, however, present in some important and influential foundational works, e.g., in the tradition of Schwinger.³⁷ As we are interested in describing QM in its relation with agency, we will use the Heisenberg picture in the discussions of this book.

3.3.2 *Quantum State Update: The Born Rule and When to Use It*

We have already alluded to the fact that the issue of updating the quantum state ascribed to a system is the basis for a number of controversies and conceptual conundra. These are mostly discussed under the heading of “the measurement problem”. Based on our discussion of the Schrödinger vs. the Heisenberg picture above, we provide the following, decidedly opinionated, view on the dynamics of quantum state update.

3.3.2.1 Quantum Experiments

Quantum mechanics is, arguably, our most successful experimental science. Many important quantum experiments were, and are to this day, “table-top experi-

³⁶ She might also change her model, e.g., the system’s Hamiltonian, in reaction to new information; agents who use QM skilfully normally know what to do. See Sect. 4.4.1 for a discussion of this case.

³⁷ See, e.g., Schwinger (2001).

ments”, which—like the Stern-Gerlach-experiment—can be built and controlled in a reasonably compact space.³⁸ As the systems studied are typically small, their characteristics are generally not directly observable in the individual case, but are only ascertainable through statistical analysis of many runs of an experiment that amplifies small-scale effects. (This is not at all uncommon in other areas of science.) A common and useful way of thinking about a quantum experiment is, then, as follows: there is a source of quantum systems, S , that prepares the systems in question in an identical way and independently of one another (identically and independently distributed, i.i.d.). The systems then undergo some interaction I , and subsequently, some suitable (amplified) outcome of the interaction is detected, D . In the Stern-Gerlach experiment, S is the oven, I is the inhomogeneous magnetic field in the vacuum tube, and D is the glass plate. The experimenter is, of course, in control of all three aspects of the experiment—the specific combination of S , I , and D would not normally occur in nature without suitable, often costly, interventions by agents doing experiments (e.g., assembling everything and securing proper conditions through shielding, from closing the door to keep the cat out to cooling, magnetic shielding, etc.). But once the system is set up in a stable way—once the equipment is calibrated, as one says—the main experimental intervention is to control the interaction, I . In the Stern-Gerlach experiment, this means, e.g., to switch the inhomogeneous field on or off, to vary its strength, or to turn the apparatus around its axis. The detection part D of an experiment gives rise to fundamental questions regarding its status.

When Is an Experimental Outcome Realised? Before an experiment takes place, the future is generally open with respect to its outcome: there are multiple different possible outcomes, whose probabilities are given by the Born rule to be discussed shortly. When the experiment is finished, one of these outcomes has been realised. The question of *when* an outcome has been realised, *when* a measurement has taken place—*by which event* the future has been settled—has been discussed controversially since the 1920s. Some consider it a non-question, others one of the deep riddles of nature. In the lab, at any rate, there is no controversy. Detectors are modelled with respect to their efficiency, built, and tested, and detector clicks are registered and results based on these data are published. The processes going on in detectors—e.g., in photomultipliers—are often understood down to the atomic level. This is reassuring when it comes to modelling the behaviour of the detectors, but it has also given rise to the following thought: if detection is due to some interaction at the atomic scale, why not describe it via QM as well? But then, the target of analysis shifts—the interaction in the detector is *not yet* the settling of

³⁸ This holds for many of the experiments that provided crucial confirmation of the emerging theory, e.g., the photoelectric effect, the double-slit experiment, or atomic spectroscopy, as well as for modern experiments, e.g., in quantum optics, such as Bell tests and quantum teleportation. Of course there are notable exceptions—think of accelerators, neutrino detectors, or gravitational wave detectors, not to speak of nuclear weapons. With suitable adjustments, however, the conceptual framework we describe in the main text can also be applied to these cases.

the open future, but only the entangling of the initial experimental system with a part of the detector (not a measurement, but a so-called pre-measurement). Settling thus must take place in some larger system. But that system can again be treated via QM, surely—and where will that stop? Perhaps only when somebody becomes consciously aware of a measurement result? Or not even then? Again the literature on this topic is huge. Terminologically, the issue is often referred to as finding the place of the “Heisenberg cut” that divides the world into quantum-mechanically describable events that are not settled, and a “classical” world of settled outcomes.

There are two initial answers to the question of when a measurement result has been settled, or, equivalently, where the Heisenberg cut should be placed. The first is a rule expressed succinctly by John Bell: “When in doubt, enlarge the quantum system” (Bell, 1987, 189). And the second answer is that most of the time, there is no doubt. Very soon, the systems involved are large enough, and the environments noisy enough, that an outcome of a measurement has been irreversibly settled. At that point, the future is no longer open with respect to the outcome. We may not always know when exactly that settling happens, but we have a large range of sufficient conditions: detector clicks and technical registration of a result as data, human or animal perception of an outcome, the production of a neutrino, or other, more experiment-specific criteria. Our fundamental answer to our question—the common thread that runs through all these criteria—is provided by the notion of irreversibility, as we will discuss in Sect. 3.3.2.3 below.

The Formal Description of Experiments Experiments can be as complicated as one wishes, e.g., involving repeated interventions of the experimenter depending on intermediate outcomes. Here we focus on basic experiments, which work with an ensemble of identically prepared systems and which are over upon detection.³⁹ For the mathematical description of such an experiment, ideally (as in textbook problems, where all the laborious setting up and calibrating is taken for granted) one starts out with a large number of quantum systems all prepared in the same state. These will be described either by a pure state ket, $|\psi\rangle$, or by a density operator, ρ , on a given Hilbert space $\mathcal{H}$ of dimension N . At the final stage of the experiment, one will, for each individual run, have one of the possible detection events that may be registered by a human observer, or logged in some electronic form. (Quite often, such a record will not resolve individual runs maximally finely, but provide some statistical aggregation of the runs, as in the silver deposited on the glass plate in the Stern-Gerlach experiment.) The empirically ascertained statistics of the detections can then be compared with the statistical predictions based on the assumed properties of source and interaction, which are given by the Born rule.

³⁹ This restriction is removed in more advanced experiments, e.g., in the context of quantum information, which will be discussed in Sects. 3.4.4 and 5.4. Several modern experimental techniques allow for the manipulation of *single* quantum systems, and non-destructive measurements on such systems are often possible.

3.3.2.2 The Born Rule

In the development of QM, it was found that any observable quantity of a physical system—for example, a particle’s position or momentum—is described by a Hermitian operator $\hat{A}(t)$ on $\mathcal{H}$, and that the possible outcomes of an experiment measuring that quantity correspond exactly to the eigenvalues of $\hat{A}(t)$. Consequently, such operators are called *observables*. Following Bohr, one can argue that such operators represent whole experimental set-ups.⁴⁰ In the simplest case, which we will assume here for the sake of clarity, $\hat{A}(t)$ has N different (real) eigenvalues $a_1, \dots, a_N$, and these correspond one-to-one to an orthonormal basis of $\mathcal{H}$, $|a_1, t\rangle, \dots, |a_N, t\rangle$ with $\langle a_i, t | a_j, t \rangle = \delta_{ij}$, that consists of the corresponding eigenvectors of $\hat{A}(t)$, so that $\hat{A}(t)|a_i, t\rangle = a_i|a_i, t\rangle$ ($i = 1, \dots, N$).⁴¹

On this approach to modelling quantum experiments, the probability of registering result a_i at the time t , for a quantum system in a pure state $|\psi\rangle$, is

$$\Pr(a_i, t) = |\langle a_i, t | \psi \rangle|^2. \quad (3.43)$$

This is the so-called *Born rule*, named after Max Born. The time parameter that appears in (3.43) reminds us that the probability depends on the time of the measurement. In the Heisenberg picture, this time dependence is inherited from the time dependence of the observable $\hat{A}(t)$, which, in turn, is a function of the dynamical variables of the system that is measured at time t . In terms of the expansion coefficients of Eq. (3.41), we can also write $\Pr(a_i, t) = |c_i(t)|^2 = c_i^*(t)c_i(t)$. For brevity, we will suppress the time index in the following. We have $\Pr(a_i) \geq 0$ by construction, and it is easy to check that these numbers in fact sum to one:

$$\sum_{i=1}^N \Pr(a_i) = \sum_{i=1}^N \langle \psi | a_i \rangle \langle a_i | \psi \rangle = \langle \psi | \left(\sum_{i=1}^N |a_i\rangle \langle a_i| \right) | \psi \rangle = \langle \psi | \hat{1} | \psi \rangle = \|\psi\|^2 = 1. \quad (3.44)$$

For systems to which a state is ascribed via a density operator ρ , the Born rule takes the form

$$\Pr(a_i) = \text{tr}(\hat{P}_{|a_i\rangle} \rho), \quad (3.45)$$

where $\text{tr}(\dots)$ is the trace, defined in Eq. (3.24) above.

⁴⁰ In modern applications, the framework we describe in this section is often extended, for example, to account for imperfect measurements or for multi-round experiments and protocols. Appropriate frameworks are then, respectively, POVMs and quantum combs; see Barnett (2009) and Chiribella et al. (2009). We will discuss generalised measurements below, and we will come back to extensions of the basic framework later when needed.

⁴¹ In the case of so-called degenerate eigenvalues, an eigenspace is a two- or more-dimensional subspace of $\mathcal{H}$; one then has some freedom in choosing a basis for that subspace.

Updating the Quantum State After an outcome of an experiment has been registered, we are in a position to update the state we ascribe to our system—there is new information available via the measurement outcome, and, obviously, the updated state must be a state that is compatible with the outcome registered. In the simple, non-degenerate case of N different eigenvalues of our observable $\hat{A}$ assumed above, this amounts to replacing the description $|\psi\rangle$ by the eigenvector $|a_i\rangle$ corresponding to the registered eigenvalue a_i :

$$|\psi\rangle \rightarrow |a_i\rangle. \quad (3.46)$$

In the general case, allowing for degenerate eigenvalues, let $\hat{P}_{a_i}$ be the projector onto the eigenspace corresponding to eigenvalue a_i . Then the *projection postulate* demands that we replace our description $|\psi\rangle$ by a new normalized ket:

$$|\psi\rangle \rightarrow \frac{\hat{P}_{a_i}|\psi\rangle}{\sqrt{\langle\psi|\hat{P}_{a_i}|\psi\rangle}}. \quad (3.47)$$

A similar formula tells us what to do in case the assigned state was not pure: for a density operator ρ , the update rule again involves projection and normalisation, via

$$\rho \rightarrow \frac{\hat{P}_{a_i}\rho\hat{P}_{a_i}}{\text{tr}(\hat{P}_{a_i}\rho)}. \quad (3.48)$$

Ideal and Generalised Measurements If one knows the complete set of eigenstates and the corresponding eigenvalues of an observable $\hat{A}$, one can compute the probabilities of the different measurement outcomes (via the Born rule) and update the quantum state for any possible outcome (via the projection postulate). The mathematical quantities needed appear in the spectral decomposition (Eq. (3.22)) of the observable that is measured,

$$\hat{A} = \sum_{i=1}^N a_i \hat{P}_i. \quad (3.49)$$

It decomposes the observable into a sum of orthogonal projectors $\hat{P}_i \equiv \hat{P}_{|a_i\rangle}$, weighted by the corresponding eigenvalues a_i . (This generalises to the case of degenerate eigenvalues mentioned in Footnote 41, in which case $\hat{P}_i \equiv \hat{P}_{a_i}$ describes projectors of higher rank.) An *ideal measurement* of the observable $\hat{A}$ is often associated with the set of projection operators $\{\hat{P}_i\}$ that appear in its spectral decomposition (Eq. (3.49)). The projectors are themselves (binary) observables $\hat{P}_i$ with two eigenvalues 1 and 0 and the properties

$$\sum_{i=1}^N \hat{P}_i = \hat{1} \quad (\text{completeness}) \quad \text{and} \quad \hat{P}_i \hat{P}_j = \delta_{ij} \hat{P}_i \quad (\text{orthogonality}). \quad (3.50)$$

The orthogonality of the projectors also means that $\text{tr}(\hat{P}_i \hat{P}_j) = \delta_{ij}$.

The description of a more general type of measurement (or experiment) is given by a *positive operator-valued measure* (POVM; see Peres, 2002; Nielsen & Chuang, 2010, Ch. 2.2). A POVM replaces the set of orthogonal projectors $\{\hat{P}_i\}$ of the ideal measurement by a set of Hermitian and positive semi-definite operators $\{\hat{O}_i\}$, which again sum up to the identity.⁴² The property that the operators $\hat{O}_i$ are positive semi-definite means that their eigenvalues are either positive or zero, which guarantees that $\text{tr}(\hat{O}_i \rho) \geq 0$ with respect to any state ρ . Since the POVM elements add up to unity, so do their expectation values, which can thus be interpreted as probabilities of different measurement outcomes i . Taken together, a POVM satisfies the relations

$$\sum_{i \in I} \hat{O}_i = \hat{1} \quad \text{and} \quad \text{Pr}(i) = \text{tr}(\hat{O}_i \rho). \quad (3.51)$$

POVMs can be used to describe situations where, for example, the implementation of a measurement is noisy or otherwise imperfect. A repeated measurement of the same quantum system may then have different outcomes. Unlike an ideal measurement (Eq. (3.48)), a POVM does not specify how to update a quantum state after a measurement—that requires a separate specification. We will come back to such a situation when describing noise channels in Sect. 3.5.4. POVMs also feature in situations where the use of non-orthogonal projectors is preferable, e.g., in order to design more efficient measurement strategies. We will come back to such a situation in Chap. 4, where we show how so-called SIC-POVMs, a special type of POVMs, play an important role for the technical research agenda of QBism.

The “Collapse of the Wave Function” Equation (3.47) gives us the another opportunity to illustrate the difference between the Heisenberg and the Schrödinger pictures via a standard introductory example in quantum dynamics, the so-called spreading of the wave packet.

We already said that the update described by Eq. (3.47) is, in the Schrödinger picture, often referred to as a “collapse”. The reason for this is as follows. For quantum state descriptions that resolve the spatial behaviour of systems, one deals effectively with so-called *wave functions* that assign a time-varying complex amplitude $\psi(\vec{x}, t) = \langle \vec{x}, t | \psi \rangle$ to points in space. In this continuous case, the mathematics is more involved, but the Born rule, applied to the position observable, simply says that the *probability density* at time t for finding the system at $\vec{x}$ is given by the modulus squared of the amplitude, $|\psi(\vec{x}, t)|^2$. In the case of a free particle of mass m ,⁴³ the dynamical variables, which are the basic observables of the system, are position $\hat{X}(t)$ and momentum $\hat{P}(t)$. Both of these observables satisfy an

⁴² The number of POVM elements, which are labelled by $i \in I$, is however not necessarily equal to the dimension N of the Hilbert space; $|I|$ can be smaller or larger than N .

⁴³ For simplicity we restrict the analysis to one dimension, i.e., we consider a particle that can move only in one spatial dimension.

eigenvalue equation similar to Eq. (3.20). For example, for the position observable, we have

$$\hat{X}(t)|x, t\rangle = x|x, t\rangle, \quad (3.52)$$

where x can take on any real value. The corresponding sums appearing in the state expansion formula (Eq. (3.12)) and the orthonormality relations (Eq. (3.11)) are replaced by integrals and Dirac delta distributions, respectively. For the expansion in terms of eigenstates of the position observable, we obtain

$$|\psi\rangle_0 = \int_{-\infty}^{\infty} dx \psi(x, t)|x, t\rangle, \quad (3.53)$$

which corresponds to Eq. (3.41), with $\psi(x, t) = \langle x, t|\psi\rangle_0$. Note that the discrete set of coefficients $c_i^H(t)$ in Eq. (3.41) here become a continuous set whose members are labelled with different real values of x . Accordingly, one can consider $\psi(x, t)$ as a function of x —which is the spatial *wave function* associated with the particle. The normalisation condition Eq. (3.17) then takes the form

$$\int_{-\infty}^{\infty} dx |\psi(x, t)|^2 = 1, \quad (3.54)$$

which implies that $\psi(x, t)$, at any time t , has to be a square-integrable function of x . Similarly, the orthonormality relations in Eq. (3.21) will take the form $\langle x, t|x', t\rangle = \delta(x - x')$ where the expression $\delta(x - x')$ should be read as a distribution, or a point measure (see Gustafson & Sigal, 2011).

An initial state assignment for the particle at time t_0 thus corresponds to specifying some initial values for the wave function, i.e., $\psi(x, t_0) = \langle x, t_0|\psi\rangle_0$. This assignment could reflect a situation where the particle is localized around some position x_0 , e.g., with the help of some trapping potential which is then switched off at time t_0 . The details of this function will depend on the potential, but often one chooses a Gaussian distribution with some mean value $x_0 = \langle \hat{X}(t_0) \rangle$ and some initial width, or spreading, $\delta X = \delta X(t_0) = (\langle \hat{X}^2(t_0) \rangle - \langle \hat{X}(t_0) \rangle^2)^{\frac{1}{2}}$ specifying the precision with which one can tell the position of the particle in the region near x_0 . Similarly, there will be some initial uncertainty in telling its precise momentum, quantified by the corresponding spreading $\delta P = \delta P(t_0) = (\langle \hat{P}^2(t_0) \rangle - \langle \hat{P}(t_0) \rangle^2)^{\frac{1}{2}}$. In the Heisenberg picture, it is straightforward to show⁴⁴ that, for the free particle, the spreading of the wave packet will grow in time. For the free particle of mass m , the Hamiltonian has the form

$$\hat{H}(t) = \frac{\hat{P}(t)^2}{2m}, \quad (3.55)$$

⁴⁴ A clear exposition, which we follow here, is given by Englert (2006a).

and Eq. (3.33) can easily be integrated, both for position and momentum, with the results

$$\hat{X}(t) = \hat{X}(t_0) + \frac{t}{m} \hat{P}(t_0) . \quad (3.56)$$

and

$$\hat{P}(t) = \hat{P}(t_0) . \quad (3.57)$$

These equations together describe the complete dynamical evolution of the system. They describe, as it were, what happens with the system (i.e., what changes) over time. Given our initial information about the system—for example, based on some state preparation at time t_0 as described above (“particle localized by the trapping potential at x_0 and then released at time t_0 ”)—we can calculate the expectation value of any observable at a later time $t \geq t_0$. For example, for $\hat{X}^2(t)$ one obtains

$$\begin{aligned} \hat{X}^2(t) &= \left(\hat{X}(t_0) + \frac{t}{m} \hat{P}(t_0) \right)^2 \\ &= \hat{X}^2(t_0) + \frac{t^2}{m^2} \hat{P}^2(t_0) + \frac{t}{m} \left(\hat{X}(t_0) \hat{P}(t_0) + \hat{P}(t_0) \hat{X}(t_0) \right) , \end{aligned} \quad (3.58)$$

while

$$\hat{P}^2(t) = \hat{P}^2(0) . \quad (3.59)$$

On the basis of these Eqs. (3.56)–(3.59) one readily obtains

$$\begin{aligned} \delta X(t)^2 &= \langle \hat{X}^2(t) \rangle - \langle \hat{X}(t) \rangle^2 \\ &= \delta X(t_0)^2 + \frac{t^2}{m^2} \delta P(t_0)^2 \\ &\quad + \frac{t}{m} \left(\langle \hat{X}(t_0) \hat{P}(t_0) + \hat{P}(t_0) \hat{X}(t_0) \rangle - 2 \langle \hat{X}(t_0) \rangle \langle \hat{P}(t_0) \rangle \right) . \end{aligned} \quad (3.60)$$

Equation (3.60) tells us that the spreading grows in time and, for large times, one obtains

$$\delta X(t) \simeq \frac{t}{m} \delta P(t_0) , \quad (3.61)$$

which means that our initial uncertainty about the position will grow linearly. In the Heisenberg picture, the explanation is very simple. Our uncertainty about the particle’s initial velocity, measured by $\delta P(t_0)$, naturally increases our uncertainty about the particle’s position at later times. In other words, our predictions for the particle’s position will become less precise over time. Note that in the Heisenberg

picture, the evolution of the system is completely described by the time dependence of the dynamic variables in Eqs. (3.56) and (3.57), which is to be distinguished from the agent's knowledge and her ability to predict the precise values of certain observables. The latter naturally deteriorates over time.

To update and maintain her prediction capabilities, the agent will need to measure the system after some time. Upon finding the system at x , say, she should then (roughly) replace the wave function, which may in the meantime be spread out over all of space, by a new function concentrated around x . This is the basis for speaking of Eqs. (3.47) or (3.48) as “the collapse of the wave function”, or “the collapse postulate”. However, one should not think of the “collapse” as a physical process like a sudden change in the electromagnetic field at a certain point in space. The particle's wave function does not describe a physical wave (in the sense of a “matter wave” or similar), and accordingly there is nothing that could collapse, except for the agent's predictions for where and with which velocity the particle will be found upon measurement. Furthermore, measurements of the basic observables can always have only finite precision. (A fundamental limit is given by Heisenberg's uncertainty relation, which will be discussed in Sect. 3.3.4.) Thus, an updated state assignment necessarily comes with its own uncertainties, which again leads to the agent's future loss of predictability.

In any case, if the system undergoes further interactions, one should restart calculations with a new state assignment updated according to Eqs. (3.47) or (3.48). This is *not* controversial, it underlies the common successful practice of modelling quantum experiments (recall our discussion of three consecutive Stern-Gerlach devices on Page 101 above). What *is* controversial, however, is how to understand this update conceptually and also physically.

3.3.2.3 Updating and Quantum Dynamics in the Schrödinger vs. the Heisenberg Picture

At this point we are in a position to expand on some of the remarks we made above about the Schrödinger vs. the Heisenberg picture of quantum dynamics. In the Schrödinger picture, the “collapse” replacement of Eqs. (3.47) or (3.48) amounts to suspending the unitary dynamics that governs the evolution of the quantum state according to the Schrödinger equation (Eq. (3.30)) and to posit an instantaneous collapse that, arguably, calls for a different kind of dynamics. Finding a physically convincing way of linking these different kinds of dynamics has proven difficult, to say the least. When the quantum state is evolved up until the time of the update, as is the case in the Schrödinger picture, there is a tendency of thinking about the update as *revealing* what was so about the system at that time anyway, but this idea is in tension with the merely probabilistic predictions about the update that can be made. (It is also just a small step from this way of thinking to the assumption that even unperformed experiments, which would simply be revealing what is so anyway, should have outcomes—which gets one into all sorts of trouble, as we will see.) In the Heisenberg picture, on the other hand, we describe how agents use QM to come

to grips with the world, e.g., for making predictions for the outcome of experiments or for understanding what the world is like. The mathematical modelling toolbox of QM allows a user of that theory to ascribe a quantum state to a system initially at a certain time, and to update the quantum state ascription at later times. On this view, it is perfectly reasonable to assume that an agent who had ascribed the quantum state $|\psi\rangle$ at the beginning of an experiment, and who had not modified that ascription during the system's interaction in the course of the experiment, should after the experiment update her state assignment to align with the experiment's detected outcome, in accord with Eqs. (3.47) or (3.48).

As we showed above, for the expansion coefficients of Eq. (3.40) it makes no difference which picture we choose, and therefore the probabilistic predictions for the time of detection are the same. This empirical equivalence, in our view, puts the Heisenberg picture approach at a strong conceptual advantage: When states are assigned by agents, it is almost trivial that an agent who has done an experiment should update her state assignment in correspondence with the outcome. On both pictures, however, the question remains which physical processes are relevantly like experiments, i.e., when an update is called for, and what in the dynamics of these processes accounts for that. When are agents *well advised* in updating their state ascription? In brief, we follow Englert (2013) in holding that it is irreversibility that justifies state updates.

“Measurement” and Irreversibility The issue of quantum state update is often discussed under the heading of “measurement”. But the common focus on measurement, and even a broader focus on experiments, appears too narrow, as argued, e.g., by Bell (1990, 34): “If the theory is to apply to anything but highly idealised laboratory operations, are we not obliged to admit that more or less ‘measurement-like’ processes are going on more or less all the time, more or less everywhere?” Bell is certainly right: It is simply a fact that quantum state updates are used in many empirically well-confirmed calculations in which nothing that one could reasonably call measurement or experiment is under consideration. Focusing on what is common in all these cases will help to steer clear of (in our view) unhelpful considerations of the role of the mind or consciousness in the measurement process,⁴⁵ and will also avoid too much emphasis on a “classical” measurement apparatus.⁴⁶

What are the defining characteristics of cases in which a quantum state update is called for? When should agents who use QM as a tool—whether for prediction, for understanding, or for engineering—update their quantum state assignments? Which situations are relevantly measurement- or experiment-like? It may not be possible

⁴⁵ See Lockwood (1996) for discussion of “many minds” approaches that links the occurrence of measurement-like processes to consciousness.

⁴⁶ Such emphasis was characteristic of the approach to QM developed in Copenhagen by Bohr and others; see, for example, Bohr (1934). We agree that once a higher ontological level in the sense of Sect. 2.2.3 is affected, a measurement-like process has happened. But that is a sufficient, not a necessary condition, as the following example of the sun clearly demonstrates.

to answer these questions via a set of completely sharply defined necessary and jointly sufficient conditions, but answering these questions has great foundational relevance. Cases that do not involve measurement or experiment will be especially important for clarification. Once the characteristics of such cases are established, it may then again be useful to revert to considerations of experiments or experimental measurements as particularly clear examples, but it should not be important, really, whether anybody looks.

As we said, in experiments, one typically ascribes a state at the source, and one updates that description at the detection stage—but other uses of state updates are also common. For example, looking back to Chap. 2, in modelling the fusion processes of the p-p-chain that produce most of the energy in the sun, one considers individual steps in a chain of reactions leading from protons (p) via the production of deuterium nuclei (d), positrons (e^+) and electron neutrinos (ν_e) first to helium-3 nuclei (${}^3\text{He}$) and gamma rays (γ) and then on to helium-4 nuclei. The first two steps in that chain are:

$$p + p \rightarrow d + e^+ + \nu_e, \quad (3.62)$$

$$d + p \rightarrow {}^3\text{He} + \gamma. \quad (3.63)$$

Note that the protons that enter the reaction in the first step are completely thermalised, like the silver atoms coming out of the oven in the Stern-Gerlach experiment. Each of these protons has undergone zillions of (“elastic”) collisions with other protons from the solar plasma before it undergoes the weak nuclear reaction of Eq. (3.62).⁴⁷ The source in the “experiment” corresponding to Eq. (3.62) would thus be a gas, or plasma, where the velocity distribution for the protons is described, to a very good approximation, by a Maxwell-Boltzmann distribution, i.e., by a thermal distribution (Adelberger et al., 2011). In describing the scattering process quantum mechanically, one thus effectively assigns quantum states to the protons that are probabilistic mixtures of projectors onto (relative) momentum eigenstates. For specific momentum eigenstates of the incoming protons, the momenta of the outgoing reaction products, including deuterium d , positron e^+ , and the electron neutrino ν_e may be, at least for a very short time, entangled. But each of the products will go their own way, and their subsequent interactions with other particles of the solar plasma will quickly lead to a loss of that entanglement.⁴⁸ Thus for the next step of the p-p-chain described by Eq. (3.63), the deuterium nucleus that enters this reaction will again be described by a quantum state that corresponds to

⁴⁷ See Fig. 3.7 on Page 192 for illustration. In other words, the scattering cross section of this proton-capture reaction is extremely small, and it takes billions of years for a given proton to undergo such an (“inelastic”) reaction.

⁴⁸ For the neutrino that is part of the reaction products, the situation may be different, as it interacts with matter only very weakly. However, the important point is that it will never come back and participate in any of the subsequent interactions of the individual other products that participated in that reaction.

a mixture of different momentum eigenstates, drawn from the appropriate thermal distribution.

For the description of the p-p-chain, and to model the energy production and particle conversion dynamics in the sun, an update of the quantum state of the participating nuclei at each step in the chain is therefore perfectly warranted. While the detailed description of the nuclear processes is rather complicated, involving weak and strong particle-changing nuclear interactions that are described by quantum field theory, the conceptual issues regarding the formalism of measurement and quantum state update are similar to the description of inelastic (and even elastic) collisional interactions in non-relativistic gases. The main issue is the irreversibility of the processes involved. We will re-encounter this issue when revisiting the Stern-Gerlach experiment (Sect. 3.3.3), and in the discussion of noise and errors in quantum error correction (Sect. 3.5.4).

In modelling the fusion processes in the sun, there is certainly no measurement taking place in any reasonable sense of the word, no experiment is being conducted, and no classical apparatus is involved either—the conditions are, to put it mildly, unfavourable. If one wants to entertain the possibility that processes such as Eq. (3.62) could be reversed “in principle” by including a sufficiently enlarged subsystem of the solar plasma, one needs to provide more details as to what is meant by the often abused term “in principle” (see Englert, 2013). This will not only refer to issues of practical (in)feasibility, but also require a serious consideration of the physical resources.⁴⁹

The common characteristics of measurement-like cases is, then, that an irreversible change has taken place. We will thus proceed under the assumption that a quantum state update in the Heisenberg picture is called for exactly in those cases in which an irreversible change in the system has occurred. That is clearly the case in experiments or measurements in which an outcome is recorded, but the above examples of scattering processes in the sun show that an actual record is not necessary, nor the involvement of a classical apparatus, let alone the presence of a conscious being.

Irreversibility It seems necessary to comment on the involvement of irreversibility in such fundamental matters. Some people balk at the notion of irreversibility and believe that it cannot play any role in foundational considerations, because, as they say, the fundamental laws of physics are invariant under time-reversal.⁵⁰ It is true that some fundamentally important laws of physics show this type of invariance, Newton’s law of motion being a prime example. But whether this

⁴⁹ For a related analysis in the context of the in-principle detectability of certain macroscopic quantum superpositions in much simpler systems, see Skotiniotis et al. (2017).

⁵⁰ John Bell, quoted above, is no exception: He includes “reversible” and “irreversible” in his list of “words which, however legitimate and necessary in application, have no place in a *formulation* with any pretension to physical precision” (Bell, 1990, 34, italics in original). His essay being brief, he gives no elaborate reason for why (ir-)reversibility ought to be excluded, saying only that these notions “defy precise definition”.

pronouncement holds true in QM is exactly what is at issue here, so that it appears to be question-begging to start from an assumption of fundamental reversibility.⁵¹ At any rate, people, and especially physicists in the lab, are very good at distinguishing reversible from irreversible processes. People also know (and are often reminded by forensic experts in crime stories on TV) that reversibility may often hold at some scale, but fail at a smaller scale. Opening a door may appear to be reversible, but the pattern of dust before a door that was closed for a week is different from that before a door that was opened and then closed a minute ago. What about reversibility at the atomic or subatomic, smallest scale?

It is important to be clear that the notion of (ir)reversibility at play in such a discussion must be the strict one familiar from considerations of microphysical reversibility in thermodynamics and statistical mechanics. In a process that is reversible in this fundamental sense, it is possible to bring the system back into the *exact same microstate* in which it started. When one uses the analogy of processes leaving or eliminating records, one needs to take into account that eliminating a record in the relevant sense does not mean simply to destroy the record (e.g., to burn it), but to get the ink back into the pen from the page into which it has diffused and to undo the frictional creation of heat during the writing process, or to reverse all physical correlates of the perception of an outcome in an observer's body. Compared to the problem of reversing the nuclear reaction described in Eq. (3.62), which involves, among other things, bringing back the neutrino to its reaction "cohort," getting the ink back into the pen would be an easy task. Note that, after the reaction in Eq. (3.62), the other particles of the neutrino's cohort go their own ways and will quickly get involved in (elastic or inelastic, annihilating) interactions with other particles of the solar plasma. Specifically, the e^+ of Eq. (3.62) will be annihilated by collisions with ambient electrons, and the d will participate in a strong nuclear reaction process described in Eq. (3.63) within a time scale of a second. So that reaction will also have to be "undone" if it takes longer than a second to get the neutrino back. The image of Humpty Dumpty, the fallen egg that cannot be put back together again, provides an adequate illustration for all of this.⁵²

In summary, our position is the following: It is agents that perform quantum state updates, and they do it successfully when they are warranted to do so by what has happened in nature. The crucial point is that an irreversible process must have occurred—no matter whether that process was observed or not, and no matter whether a record was left, or whether a classical apparatus was involved or not. Thus, the formation of a deuterium nucleus in the core of the sun is irreversible in this sense, even though the plasma in the sun's core can hardly be called a classical apparatus, and nobody could possibly be recording what is happening.

⁵¹ The mathematical fact that the Schrödinger picture time evolution operator $\hat{U}(t_0, t_1)$ has an inverse, as it is unitary, is not a good argument here, as that inverse is not physically meaningful in all cases. See Englert (2013) for discussion.

⁵² See the "Humpty Dumpty" papers by Englert et al. (1988), Schwinger et al. (1988), and Scully et al. (1989).

At the other end of the energy scale, a silver atom traversing a Stern-Gerlach apparatus undergoes an irreversible change of spin alignment when entering the fringe field of the magnet, as will be discussed below (Sect. 3.3.3). But certainly not all quantum interactions are irreversible—there are also many truly reversible processes at the quantum scale. In what follows we will discuss the case of a photon traversing a Mach-Zehnder interferometer: Its dynamics before it is detected is reversible. (Detection, however, is via absorption, whereby the photon is destroyed.) The different theoretical treatment of the reversible Mach-Zehnder dynamics vs. the irreversible Stern-Gerlach dynamics is underscored by observable effects, as detailed in the next section.

3.3.3 *The Stern-Gerlach Experiment vs. the Mach-Zehnder Interferometer*

Here we illustrate the connection between successful quantum state updates and irreversibility via two experiments, the Stern-Gerlach experiment already discussed at length above (Sect. 3.1.3), and the Mach-Zehnder interferometer, to be introduced below.

The Stern-Gerlach experiment, as we already said, perfectly fits the S - I - D scheme of a quantum experiment: the oven (source S) emits silver atoms in a state whose spin component is maximally mixed, which is described by a suitable density operator ρ . The interaction with the magnetic field inside the vacuum tube, I , then influences the atom's spin and successively gives rise to a change in momentum that deflects the atoms off the beam's central axis in a direction parallel or antiparallel to the field gradient. The glass plate acts as detector D and registers the amplified effect of that momentum change (and thus, of the spin alignment) via the place at which an atom hits the plate, some distance away from the centre of the beam. As it takes many silver atoms to produce a detectable spot on the plate, the output registered in the original experiment must be understood as an aggregate result of a huge number of individual runs of the basic one-atom experiment, but it is that one-atom experiment that the formalism of QM describes.

As we have by now introduced the necessary mathematical machinery, we can give a more formal description of the Stern-Gerlach experiment, thereby highlighting the issue of state-update.

3.3.3.1 Mathematical Description of the Stern-Gerlach Experiment

A two-dimensional Hilbert space, $\mathcal{H}$, is adequate for describing the spin of a neutral silver atom. In analysing the experiment in this way, we treat the motion of the atom as classical, which is well justified. We are also well justified in taking the

spin states of the atoms coming out of the oven to be maximally mixed, so that we can initially ascribe the density operator corresponding to a maximally mixed spin state, $\rho = \frac{1}{2}\hat{1}$, to each atom entering the vacuum tube. The question is how and when to update the state ascribed to such an atom, and based on which physical processes.

We said above that the analysis of sequential Stern-Gerlach devices gives us strong reasons to ascribe the spin state $|\uparrow, z\rangle$ to an atom coming out of the upper channel of a z -oriented device, and $|\downarrow, z\rangle$ for an atom from the lower channel. These spin states are eigenkets (eigenvectors) of a Hermitian operator $\hat{S}_z$ on the two-dimensional Hilbert space $\mathcal{H}$; in fact, they form an orthonormal basis of that space. As eigenvalues, we have⁵³

$$\hat{S}_z|\uparrow, z\rangle = \frac{\hbar}{2}|\uparrow, z\rangle, \quad \hat{S}_z|\downarrow, z\rangle = -\frac{\hbar}{2}|\downarrow, z\rangle. \quad (3.64)$$

In terms of projectors onto the subspaces of $\mathcal{H}$ spanned by these eigenkets, given that they form an orthonormal basis, their sum must be the identity operator (see Eq. (3.15)):

$$\hat{P}_{|\uparrow, z\rangle} + \hat{P}_{|\downarrow, z\rangle} = \hat{1}. \quad (3.65)$$

One can also check that the operator $\hat{S}_z$ is proportional to their difference:

$$\hat{S}_z = \frac{\hbar}{2} \left(\hat{P}_{|\uparrow, z\rangle} - \hat{P}_{|\downarrow, z\rangle} \right). \quad (3.66)$$

Spin alignment can be measured along any direction in space; nothing is special about the z -direction. Thus, the relations expressed in Eqs. (3.64)–(3.66) hold for any spatial direction $\vec{n}$ with operator $\hat{S}_{\vec{n}}$ and respective eigenvectors $|\uparrow, \vec{n}\rangle$ and $|\downarrow, \vec{n}\rangle$. The corresponding spin observable

$$\hat{S}_{\vec{n}} = \vec{n} \cdot \hat{\vec{S}} \quad (3.67)$$

is a linear combination of the three Pauli operators $\hat{S}_x$, $\hat{S}_y$, $\hat{S}_z$ (see Eq. (3.72)), which correspond to projections of $\hat{\vec{S}}$ along the three Cartesian coordinate axes.

The eigenvectors of $\hat{S}_{\vec{n}}$ can be written as linear combinations (with possibly complex coefficients) of any two vectors that span our two-dimensional Hilbert space, e.g., $|\uparrow, z\rangle$ and $|\downarrow, z\rangle$. There is some freedom in assigning a phase; the common choice results in the following:

⁵³ The occurrence of the factor $\hbar/2$ is the historical reason for calling particles with this behaviour “spin one-half particles”. These include silver atoms, but also electrons, protons, and neutrons.

$$|\uparrow, x\rangle = \frac{1}{\sqrt{2}} (|\uparrow, z\rangle + |\downarrow, z\rangle), \quad |\downarrow, x\rangle = \frac{1}{\sqrt{2}} (|\uparrow, z\rangle - |\downarrow, z\rangle), \quad (3.68)$$

$$|\uparrow, y\rangle = \frac{1}{\sqrt{2}} (|\uparrow, z\rangle + i|\downarrow, z\rangle), \quad |\downarrow, y\rangle = \frac{1}{\sqrt{2}} (|\uparrow, z\rangle - i|\downarrow, z\rangle). \quad (3.69)$$

If one represents the kets as column vectors in $\mathbb{C}^2$, a typical choice is

$$|\uparrow, z\rangle = \begin{pmatrix} 1 \\ 0 \end{pmatrix}, \quad |\downarrow, z\rangle = \begin{pmatrix} 0 \\ 1 \end{pmatrix}. \quad (3.70)$$

Hence:

$$\begin{aligned} |\uparrow, x\rangle &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \quad |\downarrow, x\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -1 \end{pmatrix}; \quad |\uparrow, y\rangle = \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ i \end{pmatrix}, \\ |\downarrow, y\rangle &= \frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ -i \end{pmatrix}. \end{aligned} \quad (3.71)$$

In this representation, the components of the spin vector operator are, up to the factor $\hbar/2$, the so-called Pauli matrices σ_x , σ_y , and σ_z :

$$\hat{S} = (\hat{S}_x, \hat{S}_y, \hat{S}_z) = \frac{\hbar}{2} (\sigma_x, \sigma_y, \sigma_z) \equiv \frac{\hbar}{2} \left(\begin{pmatrix} 0 & 1 \\ 1 & 0 \end{pmatrix}, \begin{pmatrix} 0 & -i \\ i & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} \right). \quad (3.72)$$

To see how the above formal description of state update meshes with what we know about the Stern-Gerlach experiment, we can first check which update is prescribed for an atom from the oven (mixed state $\rho = \frac{1}{2}\hat{1}$) that comes out in the upper channel of a z -oriented device, with corresponding projection operator $\hat{P}_{|\uparrow, z\rangle}$: According to Eq. (3.48) and using idempotence of the projection operator ($\hat{P}_{|\uparrow, z\rangle}^2 = \hat{P}_{|\uparrow, z\rangle}$), we have

$$\rho \rightarrow \frac{\hat{P}_{|\uparrow, z\rangle} \rho \hat{P}_{|\uparrow, z\rangle}}{\text{tr}(\hat{P}_{|\uparrow, z\rangle} \rho)} = \frac{\frac{1}{2} \hat{P}_{|\uparrow, z\rangle} \hat{1} \hat{P}_{|\uparrow, z\rangle}}{\text{tr}(\frac{1}{2} \hat{P}_{|\uparrow, z\rangle} \hat{1})} = \frac{\hat{P}_{|\uparrow, z\rangle}}{\text{tr}(\hat{P}_{|\uparrow, z\rangle})} = \hat{P}_{|\uparrow, z\rangle} = |\uparrow, z\rangle \langle \uparrow, z|. \quad (3.73)$$

So in fact, we should update such that we replace the mixed state ρ by the pure state $|\uparrow, z\rangle$. This is in agreement with the description of the empirical facts about the experiment that we gave in Sect. 3.1.3. Note that the probability for an atom from the oven to land in the upper channel of our z -oriented device according to the

Born rule (Eq. (3.45)) is one half, again in conformity with what we know about the experiment:

$$\Pr(\uparrow, z) = \text{tr}(\hat{P}_{|\uparrow, z\rangle} \rho) = \frac{1}{2} \text{tr}(\hat{P}_{|\uparrow, z\rangle} \hat{1}) = \frac{1}{2} \text{tr}(\hat{P}_{|\uparrow, z\rangle}) = \frac{1}{2}. \quad (3.74)$$

One can also check immediately that given an atom in the (newly assigned) state $|\uparrow, z\rangle$, the probability of finding it in the upper channel of a further z -oriented Stern-Gerlach device is one (see Eq. (3.43)), and the update will reproduce the state (Eq. (3.47)):

$$\begin{aligned} \Pr(\uparrow, z) &= |\langle \uparrow, z | \uparrow, z \rangle|^2 = 1^2 = 1, \\ |\uparrow, z\rangle &\rightarrow \frac{\hat{P}_{|\uparrow, z\rangle} |\uparrow, z\rangle}{\sqrt{\langle \uparrow, z | \hat{P}_{|\uparrow, z\rangle} | \uparrow, z \rangle}} = |\uparrow, z\rangle. \end{aligned} \quad (3.75)$$

This is the situation depicted in Fig. 3.2a. Similarly, the situation depicted in Fig. 3.2b, with the second Stern-Gerlach device being x -oriented, is accounted for by the corresponding mathematical expressions

$$\begin{aligned} \Pr(\uparrow, x) &= |\langle \uparrow, x | \uparrow, z \rangle|^2 = \left(\frac{1}{\sqrt{2}} \right)^2 = \frac{1}{2}, \\ |\uparrow, z\rangle &\rightarrow \frac{\hat{P}_{|\uparrow, x\rangle} |\uparrow, z\rangle}{\sqrt{\langle \uparrow, z | \hat{P}_{|\uparrow, x\rangle} | \uparrow, z \rangle}} = |\uparrow, x\rangle. \end{aligned} \quad (3.76)$$

and, analogously, $\Pr(\downarrow, x) = 1/2$ with the projection into $|\downarrow, x\rangle$.

3.3.3.2 Qubits

The mathematical description just introduced provides a useful background for establishing some of the basic terminology of quantum information theory. The unit of quantum information, the quantum bit (or qubit), is the communication value of a quantum system with a two-dimensional Hilbert space $\mathcal{H}$, such as the spin-1/2 particles abstractly described above. In other words, a qubit is the amount of information that can be conveyed, in a communication scenario, by exchanging a two-state quantum system. Whereas a classical bit can have two possible values, 0 and 1, representing a single, binary classical random variable (e.g., representing the two sides of a coin), a quantum bit has two possible eigenvalues with respect to *any choice* of a spin-1/2 observable $\hat{S}_{\vec{n}}$ as described in Eqs. (3.67) and (3.72).

Another way of expressing this is that the general state $|\psi\rangle \in \mathcal{H}$ of a qubit can be written in the form

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle, \quad (3.77)$$

where $|0\rangle$ and $|1\rangle$ are two basis states of $\mathcal{H}$ and $\alpha, \beta \in \mathbb{C}$, $|\alpha|^2 + |\beta|^2 = 1$. One typically uses this binary notation for the basis vectors of the qubit, both for convenience and to underline the analogy to the classical bit. A common choice for these basis vectors is to use the eigenstates of $\sigma_z = \begin{pmatrix} 1 & 0 \\ 0 & -1 \end{pmatrix} = \frac{2}{\hbar} \hat{S}_z$, i.e.,

$$\begin{aligned} |0\rangle &\equiv |0\rangle_z \equiv |\uparrow, z\rangle, \\ |1\rangle &\equiv |1\rangle_z \equiv |\downarrow, z\rangle, \end{aligned} \quad (3.78)$$

which then satisfy the eigenvalue equations

$$\sigma_z |s\rangle = (-1)^s |s\rangle \quad (3.79)$$

for $s \in \{0, 1\}$.⁵⁴ As indicated in Eq. (3.78), one sometimes writes $|0\rangle_z$, instead of $|0\rangle$, to explicitly refer to the basis and to distinguish it, for example, from the complementary basis of states, $|0\rangle_x \equiv |\uparrow, x\rangle$ and $|1\rangle_x \equiv |\downarrow, x\rangle$, which satisfy the corresponding eigenvalue equations $\sigma_x |s\rangle_x = (-1)^s |s\rangle_x$. The specification of the state of a quantum bit thus requires not only to specify the probabilities $p_0 = |\alpha|^2$ and $p_1 = |\beta|^2 = 1 - p_0$ with which the eigenvalues 0 and 1 are obtained when measuring σ_z , but also the phases of the complex amplitudes α and β . Classically, a biased coin with probability p_0 for landing heads would show the probabilistic behaviour of the measurement of σ_z , but there is no classical analogue of the phase that influences the probabilistic behaviour of different measurements.

As we will see in Sect. 3.4.4, the above qubit notation straightforwardly generalises to systems with more than one qubit.

3.3.3.3 Adding a Magnetic Field

Returning to the Stern-Gerlach experiment, more complicated experimental arrangements involve the application of magnetic control fields on the beam of atoms in the space between consecutive Stern-Gerlach devices. If the field is homogeneous then there is no net magnetic force that would depend on the internal spin of the atoms, so that there is no dispersion of the atomic beam. We can then treat the position and the velocity of the atoms as classical variables: an atom's position changes linearly with time in proportion to its velocity.

⁵⁴ One sometime refers, a little sloppily, to the binary variable s in Eq. (3.79) as the “eigenvalue of σ_z ”. More precisely, of course, $s = 0, 1$ denote the two eigenvalues of the projector $(\hat{1} + \sigma_z)/2$.

The dynamic effect of a homogeneous control field on the atom is described by the Heisenberg equation of motion (Eq. (3.33)), which then takes the form

$$\frac{d}{dt} \hat{S}(t) = \frac{1}{i\hbar} [\hat{S}(t), \hat{H}(t)], \quad (3.80)$$

with the Hamiltonian

$$\hat{H}(t) = -\gamma \vec{B}(t) \cdot \hat{S}(t), \quad (3.81)$$

where γ is a constant, the effective gyromagnetic ratio for the silver atom (derived from that of the electron), and $\vec{B}(t)$ denotes the homogeneous magnetic control field. Note that in our case $\gamma < 0$. In the Heisenberg picture, the time dependence of $\vec{B}(t)$ describes an external (parametric) time dependence, which is controlled or imposed by the experimenter, while $\hat{S}(t)$ describes an internal (dynamical) time dependence. Here we will assume that $\vec{B}(t) = \vec{B}$ is time independent. If we choose $\vec{B} = B\vec{e}_y$ to be oriented along the y -axis, for example, then Eq. (3.81) simplifies to

$$\hat{H}(t) = -\gamma B \hat{S}_y(t) = |\gamma| B \hat{S}_y(t). \quad (3.82)$$

With $\hat{H}(t)$ as in Eq. (3.82), Eq. (3.80) can be easily integrated, yielding the solution

$$\hat{S}_x(t) = \cos \omega t \hat{S}_x(0) + \sin \omega t \hat{S}_z(0) \quad (3.83)$$

$$\hat{S}_y(t) = \hat{S}_y(0) \quad (3.84)$$

$$\hat{S}_z(t) = -\sin \omega t \hat{S}_x(0) + \cos \omega t \hat{S}_z(0), \quad (3.85)$$

where $\omega = |\gamma|B$. This solution describes the dynamics of the spin of a silver atom that is exposed to an external magnetic field $\vec{B}$ pointing in the y -direction and having magnitude B . The dynamics can be visualised as a rotation of the spin vector observable $\hat{S}(t)$ in the x - z plane⁵⁵ with the angular frequency ω . This dynamics is also known as Larmor precession.

Using this mathematical framework, we can explicitly model consecutive Stern-Gerlach experiments, which we discussed qualitatively in Sect. 3.1.3, but now with an additional magnetic control field between the two Stern Gerlach devices. Thus, consider an experiment as in Fig. 3.2a where a beam of atoms passes two sequential Stern-Gerlach devices between which there is a constant and homogeneous magnetic field $\vec{B} = B\vec{e}_y$ that acts on the atoms while they traverse the region between the output of the first apparatus and the input of the second. We will assume that the

⁵⁵ If one feels uncomfortable with this picture, one can also consider the rotation of the corresponding expectation value $\langle \hat{S} \rangle$ of the vector observable $\hat{S}$, which (up to a factor of $\hbar/2$) is also referred to as the *Bloch vector* in quantum optics.

first Stern-Gerlach device is aligned in the z -direction, while the alignment of the second device can be along any direction $\vec{n}$.

We can now use Eq. (3.80) together with the basic formalism introduced in Sect. 3.2.3 to predict the outcome of such an experiment and to check whether the prediction is correct, in order to convince ourselves that the model we use to describe the experiment is adequate. A simple test would be to measure an observable for which we can predict the outcome with certainty, i.e., for which one possible outcome has probability one—which implies that all other possible outcomes must have probability zero. If the measured outcome differs from the one predicted with certainty, then there must be something wrong with the experiment or with our description of it. Accordingly, we may have to improve the experiment or revise our model.⁵⁶ For example, it might be the case that the magnetic field between the two Stern Gerlach devices is not perfectly homogeneous and there is a spatial variation of the magnetic field along the direction of propagation. This would give rise to an effective time dependence of the magnetic field as “seen” by the atomic spin while the atom moves between the two subsequent Stern-Gerlach devices.⁵⁷

Consider now the idealised set-up described above: Atoms that leave the first Stern-Gerlach device at time $t_0 = 0$ traverse a homogeneous magnetic field for a time $t = \frac{\pi}{2|\gamma|B} = \frac{\pi}{2\omega}$ until they enter the second Stern-Gerlach device, oriented along the direction $\vec{n} = \vec{e}_x$. According to our model of the interaction of the atoms with the magnetic field in the intermediate region, which is represented by the Hamiltonian (3.82), the spin vector of the atoms before they enter the second apparatus can be written as in Eq. (3.83), with $t = \frac{\pi}{2\omega}$, or $\omega t = \pi/2$:

$$\hat{S}_x(t = \frac{\pi}{2\omega}) = \hat{S}_z(0), \quad (3.86)$$

$$\hat{S}_y(t = \frac{\pi}{2\omega}) = \hat{S}_y(0), \quad (3.87)$$

$$\hat{S}_z(t = \frac{\pi}{2\omega}) = -\hat{S}_x(0). \quad (3.88)$$

Atoms that leave the first, z -oriented Stern-Gerlach device in the upper channel are described by the state $|\uparrow, z\rangle$. The corresponding expectation values of the spin in z and x directions are $\langle \hat{S}_z(0) \rangle = 1$ and $\langle \hat{S}_x(0) \rangle = 0$, respectively. Using Eq. (3.86) one thus obtains $\langle \hat{S}_x(t = \pi/(2\omega)) \rangle = 1$. The prediction for the outcome of a spin measurement along the x -direction at time $t = \pi/(2\omega)$ is then, with certainty, that the atoms will leave the second Stern-Gerlach device in the upper channel.

⁵⁶ See Sect. 4.4.1 for a discussion of the different types of model updates that may be involved here.

⁵⁷ If the atom moves with velocity v_0 in the y direction, a spatial variation of the strength of the magnetic field $\vec{B}(y) = B(y)\vec{e}_y$ along its path between the two Stern-Gerlach devices leads to an effective (parametric) time dependence of the form $\hat{H}(t) = -\gamma B(vt)\hat{S}_y$.

If we run such an experiment and find an atom in the lower channel of the second Stern-Gerlach device, in disagreement with our prediction, then we would need to check the set-up and possibly revise our model of the experiment. Hacking (1983) vividly describes that experiments have, as it were, a life of their own. A priori, it is not settled what needs to be revised if things go wrong and our expectations are violated. If we trust that the Stern-Gerlach devices used for the preparation and measurement of the atomic spin function properly and are well calibrated, then a natural candidate for revision would be the Hamiltonian that describes the atomic interaction with the B -field, e.g., by assigning a different value to B and thus to the frequency $\omega \propto B$ that describes the rotation of the spin vector (Eqs. (3.83)–(3.85)).

If one is confident about the magnetic field, then one may look more closely into the set-up of the experiment. Another candidate for revision could be the time $\tau = d/v_0$ that it takes for an atom to traverse the distance d between the two devices. That time depends on the atom's velocity v_0 . The effect of a different value of v_0 on the rotation of the spin vector (Eqs. 3.83–(3.85)) is similar to the effect of a different value B of the magnetic field that acts on the atom along the way. Now the velocity of an atom that comes out of the oven is itself a random variable v with some probability distribution $w(v)$ (close to a Maxwell-Boltzmann-distribution). That distribution will be peaked around some velocity $v = v_0$, but still have a certain width, say, $\Delta(v_0)$. Our model of the velocity distribution of the atoms in the beam will then influence our prediction of the outcome of a spin measurement, that is, the probabilities such as $\Pr(\uparrow, x, t = \pi/(2\omega))$ with which an atom will be detected in one of the output channels of the second device. Similarly, the precise directional alignments of the two Stern-Gerlach devices could be candidates for revision, and there are further options.

Once we have successfully improved the experimental set-up—or revised the model to better capture the experimental situation—such that various tests are passed and we accept the model of the experiment as accurate, we can use the set-up for various applications. One application of the above set-up could be the generation of random bits, a topic that will be addressed in Sect. 3.5.

Our descriptions of different Stern-Gerlach experiments, including extensions that add some dynamical control on the spin of the atoms, illustrate the use of the Heisenberg picture and its suitability from an agency perspective. Throughout the description of the experiment, we hold on to the quantum state ascription $|\uparrow, z\rangle$, which tells the story of how the system was prepared via the first Stern-Gerlach device. While the system interacts with a magnetic field, nothing in that story changes. Based on this ascription, we can make predictions about future measurement results, using the Heisenberg equations of motion for the observables, which describe the system's evolution over time. Our success in predicting what happens will depend on how accurately our overall description of the system and its dynamics was chosen, and on how well the experiment was actually controlled.

The main point is that our knowledge about the system during the run time of the experiment does not change. We know about its preparation (the state we assigned when we selected it), and we know about its dynamics. If our model of the preparation and the dynamics is good, we will be able to make good predictions.

This can be checked experimentally. As we pointed out, it is not predetermined what we do upon mismatch. Our reaction in such a case depends on our skilful handling of the theory. We may update our characterisation of the source, or our description of the system dynamics, but we may also just throw out several experimental runs as a fluke.

3.3.3.4 Reversibility and Irreversibility in the Stern-Gerlach Experiment

The interaction of a silver atom with a constant homogeneous magnetic field that we just described is a prime example of reversible quantum dynamics. Looking at Eqs. (3.86)–(3.88), one sees that a further interaction with a magnetic field of the same strength for a longer time t will lead to the spin directions along the x - and along the z -axis being flipped again. Consequently, an interaction for a total time of $t = 2\pi/\omega$, assuming ideal conditions, will recover the initial spin direction. The overall set-up that we describe here corresponds to a part of the more complex nuclear magnetic resonance (NMR) experiments introduced by Isidor I. Rabi and coworkers to study the magnetic moments of different atomic nuclei (Rabi et al., 1938, 1939; Hurwitz, 1992). The technique of NMR spectroscopy developed in these experiments, which involves an additional, time-dependent oscillatory magnetic field, has found widespread applications, e.g., in chemistry and material sciences, and in medical imaging. Coherent spin oscillations as described by Eq. (3.83) have been realised in various physical systems, going beyond the original atomic beam set-up used by Rabi, and have become a benchmark for single-qubit control in quantum information experiments (Vandersypen & Chuang, 2005).

In experiments involving Stern-Gerlach devices, there are thus some quantum-mechanical interactions that are perfectly reversible. As we have described above, if you take an atom in the $|\uparrow, z\rangle$ state and let it traverse a certain homogeneous magnetic field, it will continuously change its orientation and in the end (given a suitable field) align in the x -direction. This can be checked by sending it through an x -aligned Stern-Gerlach apparatus, where it will come out at the upper channel with a probability of 1, ascertaining that its state is now to be described as $|\uparrow, x\rangle$. That atom can then be subjected to a continuous full reversal of the change of orientation, e.g., by sending it through a flipped version of the initial homogeneous magnetic field, so that it is again aligned in the z -direction. This can then be checked via a third, z -aligned Stern-Gerlach apparatus.

On the other hand, there are also irreversible processes going on. The production of the image recording the outcome of a Stern-Gerlach experiment is a perfect example of this type of dynamics: The inelastic collision of a silver atom with the glass plate is as irreversible as it gets. The atom gets stuck to the plate, photons of different energies are scattered in all directions, and there is an interaction with a multitude of further uncontrolled and uncontrollable degrees of freedom of the environment. By using the state update formalism, we have implied that the physical interaction that the silver atom undergoes in the Stern-Gerlach device is irreversible upon detection, and that is certainly so. But what about times in between the source

and the screen? Where can we draw a line? When, that is, are we warranted to update the quantum state in a Stern-Gerlach experiment? Should we update at the output channels already, or can we undo the splitting of the incoming beam in some clever way? And if we cannot, what in the physics of the interaction makes it so? For answering this question, our classical intuitions are not always to be trusted.

In order to discuss the question of irreversibility in the Stern-Gerlach experiment in more detail, it will be useful, as a contrast, to first discuss a somewhat similar but fully reversible set-up, the Mach-Zehnder interferometer.

3.3.3.5 The Mach-Zehnder Interferometer

Most readers will be familiar with the famous double-slit experiment, in which the wave-particle duality of quantum objects (photons in a beam of light, or electrons, for example) can be demonstrated via the presence or absence of interference effects under different conditions: if the incident beam passes through both slits, the partial beams interfere, but the interference vanishes once there is “which way” information, e.g., from blocking one of the slits.⁵⁸ Rather than discuss the double-slit experiment, we will focus on a similar interference experiment, the Mach-Zehnder interferometer, in view of its more explicit parallels with the Stern-Gerlach experiment and its role in photonic implementations of quantum information (see, e.g., Sect. 5.6). In the Mach-Zehnder interferometer, the double slit is replaced by a beam splitter, and interference can be monitored by beam recombination, effected by a second beam splitter.

The basic structure of a Mach-Zehnder interferometer is pictured in Fig. 3.3. A light beam enters on the lower left and hits a beam splitter. The two partial beams then traverse different arms of the set-up, each containing a mirror that reflects the partial beams onto a second beam-splitter, where these partial beams are recombined. In one of the paths there is a tunable phase shifter that influences the way in which the partial beams interfere. If they arrive in phase ($\varphi = 0$), they interfere constructively, and the intensity of the horizontal output equals that of the input beam, while the intensity of the vertical output is zero. Fully out of phase ($\varphi = \pi$), there is destructive interference, with intensity zero at the horizontal output and maximum intensity at the vertical one. In between, the intensity of the horizontal beam varies in proportion to $\cos^2(\varphi/2)$. So far, the apparatus can be understood in terms of the wave nature of light as described by classical electromagnetism—but there are also truly quantum effects.

Quantum Interferometry Like the Stern-Gerlach experiment, a Mach-Zehnder interferometer can also be operated with a very dim beam, such that only a single photon is in the apparatus at any given time. In that case, there are no partial beams to interfere any more. Nonetheless, the statistics of the interference effect

⁵⁸ A nice exposition of the experiment is given in the Feynman lectures (Feynman et al., 1989, Ch. 1).

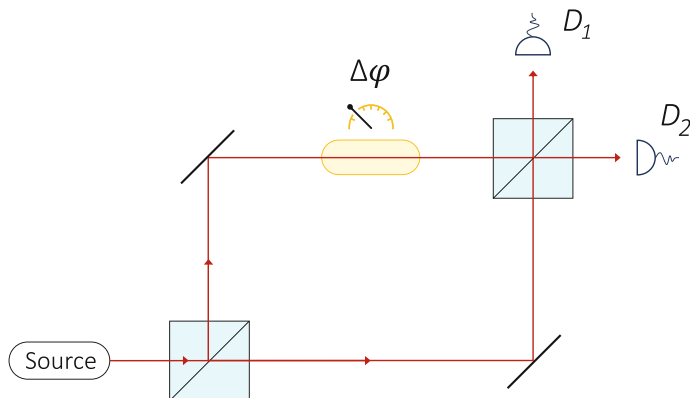


Fig. 3.3 Schematic drawing of a Mach-Zehnder interferometer. Light from a source passes a beam splitter (lower left). In one arm of the set-up there is a tuneable phase-shifter. By varying $\Delta\varphi$, interference effects can be registered via the two detectors (D_1 , D_2) after the partial beams are recombined by a second beam splitter (upper right)

persist also under these conditions: for $\varphi = 0$, all the photons come out at the horizontal output. Instead of output beam intensities, one now has to ascribe single-case probabilities for a photon passing through the interferometer to be detected at the horizontal ($\Pr(h) = \cos^2(\varphi/2)$) or vertical ($\Pr(v) = \sin^2(\varphi/2)$) output channel. We should note that such interference effects are not only possible with photons—similar interference experiments have been conducted with electrons, with neutrons (Rauch, 1998), and also with atoms (Cronin et al., 2009) and molecules (Hornberger et al., 2012).⁵⁹

Quantum interferometers have both practical and fundamental relevance. Neutron interferometry made it possible for the first time to experimentally demonstrate the fundamental 4π symmetry of fermions (Rauch et al., 1975; Werner et al., 1975). In that experiment, a magnetic field in one arm of the interferometer operated on the neutron nuclear magnetic moment in a way similar to the case of sequential Stern-Gerlach devices discussed earlier. Light interferometry has meanwhile become the basis for advanced technologies. Nowadays, arrays of Mach-Zehnder interferometers can be realized on microfabricated integrated photonic chips that are commercially available and which can be used in various quantum information applications (Wang et al., 2020),⁶⁰ including a photonic implementation

⁵⁹ Interferometers using massive particles are also known under the term *matter-wave interferometry*, highlighting the fact that these particles can be seen as individual quanta of matter fields, in a similar way as photons are the quanta of light fields.

⁶⁰ Examples given by Wang et al. (2020) include boson sampling, integrated quantum random number generation, quantum molecular simulation, quantum communication, measurement-based quantum computation, and others. Another recently proposed applications is in quantum-enhanced reinforcement learning (Saggio et al., 2021).

of Quantum Projective Simulation, which will be discussed in Sect. 5.6. The Mach-Zehnder interferometer has also been used to discuss fundamental features of quantum mechanics, e.g., in the context of quantum erasure and delayed-choice experiments, which we will discuss shortly. Such experiments play an crucial role in the discussion of quantum measurement.

An even closer analogy between the (hypothetical) Stern-Gerlach inteferometer and the Mach-Zehnder interferometer is obtained if one considers a variant of the Mach-Zehnder set-up which works with polarised beams of light. The photons then carry polarisation as an additional degree of freedom (beyond the path, or spatial mode), and the beam splitters in Fig. 3.3 are replaced by polarising beam splitters.⁶¹ This means that the direction, i.e., the arm into which an incoming photon is deflected by the polarising beam splitter, depends on the photon's polarisation state, in analogy with the Stern Gerlach set-up in which the spatial deflection of an incoming atom by the Stern-Gerlach magnet depends on its internal spin state. Similarly, the tunable phase shifter can be replaced by a polarisation rotator which changes the polarisation state of a photon, examples being a half- or quarter-wave plate. In such a set-up, the path of a photon inside the Mach-Zehnder interferometer becomes correlated with its polarisation, in analogy with a Stern-Gerlach device in which the path of an atom is correlated with its spin state. For most of the following discussion we can use the standard set-up of the Mach-Zehnder interferometer, which works with unpolarised light, but we will invoke the modified set-up whenever that is beneficial for clarity.

Quantum Coherence The phenomenon of interference that is exhibited by individual quantum systems is an expression of *quantum coherence*. The term coherence is known from classical waves and indicates a situation where the oscillations of the wave's amplitude at two different points in space have a fixed phase relation. This is, for example, the case for coherent light that is emitted by a laser, while it is not the case for the incoherent light that reaches us from the sun. In quantum mechanics one speaks, analogously, of a fixed phase relation of the different amplitudes c_i that appear in the expansion of a quantum state $|\chi\rangle = c_1|b_1\rangle + \cdots + c_N|b_N\rangle$ as in Eq. (3.12). If one writes $c_i = |c_i|e^{-i\varphi_i}$ and $c_j = |c_j|e^{-i\varphi_j}$, then the phase difference $\varphi_i - \varphi_j$ determines whether in the expectation value $\langle\hat{A}\rangle = \langle\chi|\hat{A}|\chi\rangle$ of any observable $\hat{A}$ the amplitudes c_i and c_j interfere constructively or destructively.

⁶¹ These are optical elements which deflect incoming vertically polarised light (whose electric field vector oscillates orthogonal to the plane of the figure) into the upward pointing arm, while they let horizontally polarised light (whose electric field vector oscillates in the plane of the figure) pass through into the arm that points to the right. An incoming light beam that is linearly polarised in the direction 45° is then split up into two partial beams of equal reduced intensity that leave the polarising beam splitter with the corresponding horizontal or vertical polarisations. For a very dim incoming beam, the apparatus will at most contain a single photon, whose quantum state is $|45^\circ\rangle = \frac{1}{\sqrt{2}}(|H\rangle + |V\rangle)$. Such a photon will be deflected or transmitted into one of the outgoing modes with probability one half.—The same holds for a circularly polarised beam.

Accordingly, quantum coherence is often characterised as the *capacity for interference*.⁶² It often requires a special effort and almost perfect control to prepare a quantum system such that it exhibits this capacity, i.e., to prepare it in a pure quantum state of the form Eq. (3.12). In case of imperfect control or noisy preparation, the state of the system is no longer described by a pure state but by a density operator $\rho = \sum_{m=1}^M w_m |\chi_m\rangle\langle\chi_m|$ as in Eq. (3.23), which corresponds to a probabilistic mixture of different pure states $|\chi_m\rangle = c_1^{(m)}|b_1\rangle + \cdots + c_N^{(m)}|b_N\rangle$, with $c_i^{(m)} = |c_i^{(m)}|e^{-i\varphi_i^{(m)}}$. A particular phase relation $\varphi_i^{(m)} - \varphi_j^{(m)}$ will be different for each of the states $|\chi_m\rangle$ that contribute to ρ . Accordingly, there will be a mixture of terms that interfere both constructively and destructively in $\langle\hat{A}\rangle = \text{tr}(\rho\hat{A})$, which together may wash out any interference.

Loss of Coherence and Which-way Information The degrading (or loss) of coherence is typically due to an irreversible interaction of the system with uncontrolled—and often practically untraceable—degrees of freedom of its environment. For photons, this could be random (forward) scattering interactions with molecules in the atmosphere, e.g., in the context of free-space quantum communication with satellites.⁶³ But the coherence of a quantum system can also be lost due to a reversible and controlled interaction with another quantum system that serves as a which-way detector.⁶⁴ The degree of coherence can be quantified by the visibility $\mathcal{V}$ of the fringes of the interference pattern that can be obtained. For the double-slit experiment, the pattern corresponds to a spatial variation of the detection probability on the screen, while in the Mach-Zehnder interferometer, it corresponds to an oscillation of the probability as a function of the phase φ in Fig. 3.3. It can be shown that maximal visibility ($\mathcal{V} = 1$), which is required for an ideal operation of the interferometer, implies that no information can be obtained about the path of a photon in the interferometer. This is just another expression of the wave-particle duality known from the double-slit experiment. In the case of the Mach-Zehnder interferometer, it means that no information can be acquired regarding *which way*—horizontal or vertical—a photon exits the first

⁶² As we have illustrated with the different examples given above, quantum coherence can be observed in elementary quantum systems, the simplest being a two-state system, or qubit. But it can also be observed in composite systems that are composed of two or more elementary systems. Quantum coherence in such multi-partite systems gives rise to the phenomenon of entanglement (as discussed in Sect. 3.4.3) which is a central resource in quantum information (see Sect. 3.4.4).

⁶³ We will discuss a mathematical model for qubits subject to noise in Sect. 3.5.4.1.

⁶⁴ For photons, one can think of individual, trapped, atoms that interact with the photons on their way from the source to the screen, as in Scully and Drühl (1982), such that internal quantum states of the atom get entangled with the path of the photons. Analogously, for atoms one can think of micromaser cavities that interact with atoms that pass through them and thereby drop (or absorb) a photon on their way (Scully et al., 1991). Both of these scenarios—an individual atom as a which-way detector for photons passing by, and a micromaser cavity as a which-way detector for atoms passing through—are somewhat idealized, but for the following discussion, the experimental details are not important, and we will often invoke these scenarios. A review of different types of delayed-choice gedanken experiments and their realizations has been given by Ma et al. (2016).

beam splitter and subsequently traverses the interferometer, without degrading (or completely destroying) the interference fringes. Quantitatively, it was shown by Englert (1996) that the fringe visibility sets an upper bound on the amount of which-way information (called distinguishability $\mathcal{D}$) that can be obtained from detectors placed in the two arms of the interferometer:

$$\mathcal{D}^2 + \mathcal{V}^2 \leq 1 \quad (3.89)$$

where $0 \leq \mathcal{D}, \mathcal{V} \leq 1$.⁶⁵

The Possibility of Quantum Erasure While the uncontrolled interaction of a photon with molecules in the atmosphere and the controlled interaction with a single which-way entangling quantum system have the same effect on the fringe visibility, there is an essential difference between the two. In the first case, the loss of coherence results from a practically irreversible interaction of the photon with many environmental degrees of freedom representing, e.g., the vibrational and motional states of the air molecules. In the second case, it is due to a highly controlled interaction of the photon with the internal states of, e.g., a single trapped atom. In both cases, one can say that which-way information after the interaction is stored in the state of the photon's environment. The difference has to do with the (im)possibility of *erasing* that information in the two cases. While in the first case (environment consisting of many air molecules) this is practically impossible, practical schemes of quantum erasures for the second case (environment consisting of a single trapped atom) have been devised.

This difference also plays a role in the description of a measurement in quantum mechanics. While the notion of a which-way detector, as used in the above discussion and in many original papers, suggests that the interaction of a photon with the which-way-detector system already constitutes a detection, and thus a measurement, of its state (“where did it go through?”), this need not be the case. To avoid this misunderstanding, we prefer to describe the “which-way detector” as a *which-way entangler system*: it entangles the path (quantum degree of freedom) of the photon with the internal (energy quantum) degree of freedom of the which-way-detector atom. In a full description of a measurement process, this interaction corresponds only to the first phase of entangling the state of the to-be-measured system with the pointer state of the measurement apparatus (Busch et al., 1996), a phase that is also referred to as a *pre-measurement* and which is still reversible. A full quantum measurement involves a second, irreversible phase, where the state of the pointer interacts with its environmental degrees of freedom.⁶⁶

⁶⁵ For a related discussion see also Greenberger and Yasin (1988). A recent study of relations between quantum coherence and which-way information in multi-path interferometers, which also provides further references, was given by Bagan et al. (2016).

⁶⁶ These could be, for example, the different modes of the radiation field into which the atom's excited state could decay, or, similarly, the degrees of freedom of a photodetector involved in the absorption of the photon and the subsequent amplification process. In the example of the

In a quantum measurement, the time scale of the second, irreversible step is typically much shorter than of the first step. A separation of the two phases, if it is possible, allows one to *quantum erase* the information stored in the which way entangler (i.e., the correlations between the state of the entangler and the path of the photon) in such a way as to retrieve the interference fringes (Wheeler & Zurek, 1984; Scully & Drühl, 1982). Erasure is no longer possible after the second step.⁶⁷ In “Wigner’s friend” scenarios, one integrates the perspectives of different “observers” (e.g., Wigner and his friend) and interprets the different terms of the entangled state after a pre-measurement *as if* they were the results of a real measurement. This, however, leads to inconsistencies, as demonstrated by Frauchiger and Renner (2018) and discussed by Żukowski and Markiewicz (2021).⁶⁸

The process that is described by the uncontrolled interactions of a photon with the atmosphere is indeed similar to the second main aspect of a quantum measurement, namely irreversibility, that we are interested in. It highlights the part of a measurement that *settles matters* about an individual quantum system. The suppression of unwanted, uncontrolled interactions of the system with any other degrees of freedom is indeed the main challenge for the different realisations of a quantum interferometer. It will also be the main challenge for applications of quantum technologies, including quantum memories and the quantum computer, which exploit the quantum coherence (and thus the entanglement) of larger, multi-partite quantum systems. This is a topic that we will address in the next section and revisit in our discussion of quantum information applications and quantum error correction (Sects. 3.4.4 and 3.5.4). For the Mach-Zehnder interferometer, this challenge can be met. With a view to reversibility, the upshot is that the splitting of the beam into two components by the initial beam splitter, even in the case of a dim beam consisting of single photons, is fully reversible; coherent recombination is possible. Exactly as in the double-slit experiment, however, coherence is lost once “which path” information is available.

which-way detector in Scully et al. (1991), the micromaser cavity that can store a single photon is a realisation of the pre-measurement phase. The second step, the completion of the which way *measurement*, would then consist, e.g., in the irreversible absorption of the photon by some thin-film semiconductor material that acts as a photodetector.

⁶⁷ That would require a projection of the system into a basis of highly entangled states of the pointer variable with the environmental degrees of freedom of the detector.

⁶⁸ The performance of a quantum measurement requires both phases. A well-known slogan by Peres (1978), who addresses counterfactual reasoning in the context of quantum experiments, says that “unperformed measurements have no outcomes”. Iterating on this slogan, Żukowski and Markiewicz (2021) argue that “even performed pre-measurements have no outcomes”. (Remember that the term “pre-measurement” denotes just an entangling operation between the system and the pointer, which typically leaves the reduced state of the system undetermined, i.e., in a completely mixed state.) While we disagree with the conclusions of Frauchiger and Renner (2018) regarding the general state of affairs in quantum mechanics, it is big a merit of their work to have mapped out, in a detailed multi-partite scenario improving greatly on the usual Wigner’s friend-type arguments, the implications that follow from interpreting pre-measurements as real measurements.

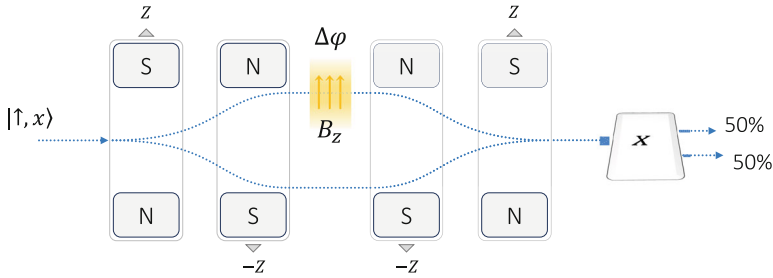


Fig. 3.4 A hypothetical Stern-Gerlach interferometer consisting of four Stern-Gerlach devices. By concatenating the devices with the orientations shown, the partial atomic beams coming out of the first device are spatially recombined. Note the structural correspondence with the Mach-Zehnder interferometer shown in Fig. 3.3; the tuneable homogeneous magnetic field in the upper channel between the two middle devices corresponds to the tuneable phase-shifter in one of the arms of the Mach-Zehnder set-up. The question is whether the atomic beams here are recombined coherently, i.e., whether interference effects can be demonstrated

It will now be instructive to consider analogies and disanalogies in the Stern-Gerlach experiment.

3.3.3.6 A Stern-Gerlach Interferometer?

The question of reversibility vs. irreversibility is already present and takes on an especially pure form in the case of the Stern-Gerlach set-up: Is a Stern-Gerlach interferometer possible, i.e., is it possible to recombine the two partial beams at the output of a Stern-Gerlach apparatus into a single beam in a coherent way? Is it possible to have a strict analogue of a Mach-Zehnder interferometer built not from optical beam splitters, but out of Stern-Gerlach devices? In such a (hypothetical) set-up, illustrated in Fig. 3.4, one would have an incoming beam of atoms in spin state $|\uparrow, x\rangle$ that is split into two sub-beams via a z -oriented Stern-Gerlach device and then subsequently recombined. If this set-up works as an interferometer, the initial pure state will be recovered, i.e., it will be correct to ascribe the spin state $|\uparrow, x\rangle$ to all atoms coming out of the set-up after the partial beams have been recombined. If, on the other hand, no coherent recombination occurs, then the output state after the spatial beam recombination should be a mixture of the spin states $|\uparrow, z\rangle$ and $|\downarrow, z\rangle$, to be described by a density operator $\rho = 1/2(|\uparrow, z\rangle\langle\uparrow, z| + |\downarrow, z\rangle\langle\downarrow, z|)$. This difference is easy to verify empirically: If we feed the output into an x -oriented Stern-Gerlach apparatus, in the first case (coherent recombination), we expect all atoms to come out at the upper channel, while in the second case (mixture), we expect a 50:50 split of the atoms in the two output channels. An additional test would be to see what happens if we add a homogeneous magnetic field (as discussed in Sect. 3.3.3.3) in one arm of the interferometer set-up. For this test, in the first (coherent) case, we would see oscillations of the detection probabilities at the output

of the final x -oriented apparatus that vary with the strength of the field B ; in the second case (mixture) there would be no effect.⁶⁹

Note that, in the first case, the interferometric recombination of the beams would constitute a successful quantum erasure of the pre-measurement (i.e., the entanglement of the atoms' trajectories in the purported interferometer with their internal spin state). It would be a reversible process and the final, real measurement with the help of an x -oriented Stern-Gerlach apparatus would just verify the successful coherent recombination. In contrast, in the second case, the purported interferometer would in fact destroy the coherent superposition of the input state $|\uparrow, x\rangle = \frac{1}{\sqrt{2}}(|\uparrow, z\rangle + |\uparrow, z\rangle)$ and turn it into a mixed state.

Based on our discussion above, the question of which description is right amounts to the question of whether the mere passage of a spin-1/2 atom through a Stern-Gerlach apparatus already is an irreversible process or not. Does the mere passage of the atom through the inhomogeneous field—without any detection at the glass plate or a similar recording of “which path” information—already constitute a measurement of spin projection or not?

The answer is, roughly speaking, “it depends, but yes”. For realistic set-ups, the process is irreversible, and thus it has all the characteristic features of a measurement of spin projection. Of course, some remarks are in order.⁷⁰ First, as we said above, the registration of an atom by collision with the glass plate at the end of the apparatus is irreversible and constitutes a measurement—directly, of position, and indirectly, thereby, of spin projection. This is generally not disputed. The glass plate is a typical “classical apparatus”, and the collision also marks the end of the atom's trajectory as a system under investigation.

Second, the question of reversibility of the interaction *before* the atom has collided with anything has been the subject of much research and debate. We already mentioned the analogy with Humpty Dumpty, the egg that could not be put back together again after his great fall, which provides the leading theme of a famous series of papers (see Footnote 52). These papers show that a successful coherent recombination of the partial beams would require control of the inhomogeneous magnetic fields in the four devices with sub-microscopic precision, which is unattainable. (By contrast, the devices involved in optical interferometry are uncritical from that perspective.)

In comparing the deflection of atoms in a Stern-Gerlach device with beam-splitting in a Mach-Zehnder interferometer, the specific physics of the set-up needs to be taken into account—e.g., the existence of a fringe field at the entrance of the region in which the inhomogeneous magnetic field acts on the incoming atoms. It

⁶⁹ The latter prediction follows from the fact that the mixture $\rho = 1/2(|\uparrow, z\rangle\langle\uparrow, z| + |\downarrow, z\rangle\langle\downarrow, z|)$ is the same in any direction. That is, $\rho = 1/2(|\uparrow, x\rangle\langle\uparrow, x| + |\downarrow, x\rangle\langle\downarrow, x|)$; see Eq. (3.90).

⁷⁰ There is also some fine print. We are here talking about the *transversal* Stern-Gerlach experiment, as it was originally conceived and executed in 1922, with the inhomogeneous field's gradient perpendicular to the beam. There is also a set-up called the “longitudinal Stern-Gerlach experiment”, in which the direction of the beam (or rather, the direction of motion of a cluster of gravitationally freely falling cold atoms) and field gradient are parallel. This set-up has different characteristics and does not amount to a measurement, being reversible (Margalit et al., 2021).

is clear that quite generally, interferometry with atoms is possible (the experiment mentioned in Footnote 70 is a case in point; see also Footnote 59). The specific question under discussion here is, however, whether the splitting induced by the magnetic field in the original (transversal) Stern-Gerlach set-up allows for coherent recombination or not.

Third, the most accurate available calculations of that problem (Paraniak & Englert, 2021) show that even assuming ideal control of the magnetic field (which, as mentioned, is unattainable), the possibility of recombination depends critically on the spatial separation of the two components. Below a threshold of around 10 diameters of the atomic wave-packet, coherent recombination could in principle be achieved; over a threshold of around 35 diameters, the process is irreversible, and a coherent recombination of the partial beams is impossible.

In a practical realisation of the Stern-Gerlach experiment as in Fig. 3.1, the threshold for recombination is exceeded by orders of magnitude, so that the experiment—without any recording of the atoms' position, and even if the atoms are all brought back to the original axis of flight—amounts to an irreversible measurement of spin projection. This is why it is adequate to replace the description of the state of an incoming atom in the set-up of Fig. 3.4, $|\uparrow, x\rangle$, by either $|\uparrow, z\rangle$ or $|\downarrow, z\rangle$ at the respective output of the apparatus. And if the two partial beams coming out of that apparatus are spatially recombined, the correct description is not $|\uparrow, x\rangle$ (which would be the case if the splitting had been fully reversed), but a mixture of $|\uparrow, z\rangle$ and $|\downarrow, z\rangle$ (which is mathematically the same as a mixture of $|\uparrow, \vec{n}\rangle$ and $|\downarrow, \vec{n}\rangle$ for any direction $\vec{n}$):⁷¹

$$\rho = \frac{1}{2}(|\uparrow, z\rangle\langle\uparrow, z| + |\downarrow, z\rangle\langle\downarrow, z|) = \frac{1}{2}\hat{1} = \frac{1}{2}(|\uparrow, \vec{n}\rangle\langle\uparrow, \vec{n}| + |\downarrow, \vec{n}\rangle\langle\downarrow, \vec{n}|). \quad (3.90)$$

The irreversible interaction of the incoming atom with the fringe field of the magnet results in one of two possible alignments, parallel or antiparallel to the field gradient. Once that interaction has taken place, a state update is warranted, and the interaction is, in effect, a spin projection measurement—no matter whether an outcome is recorded or not.

⁷¹ Remember, however, that loss of coherence is not an all-or-nothing affair, it can be quantified. As we pointed out earlier, the trade-off between coherence (the capability for interference) vs. the availability of “which path”-information is described by the Englert inequality (Eq. (3.89)). For a partial loss of coherence, the resulting density operator is described, in the simplest case, by Eq. (3.155). For the present set-up, one then has $\rho = (1 - p)|\uparrow, x\rangle\langle\uparrow, x| + p|\downarrow, x\rangle\langle\downarrow, x|$ with values of $0 \leq p \leq 1/2$ quantifying the degree of “dephasing.” The maximum visibility in this case would be $\mathcal{V} = 1 - 2p$.

3.3.4 Complementarity and the Heisenberg Uncertainty Relation

We had already mentioned the notion of complementarity above. Based on our discussion of quantum state update, we can now discuss that notion in more detail. First, we can draw out two points from our discussion of when agents are warranted to update the quantum state. One is that updating is tied to irreversible happenings: an agent's state update would be premature while the change in the system under consideration is still reversible, but it becomes warranted once that change can no longer be reversed. The second point is that the way to update the quantum state via the “collapse” rule (Eqs. (3.47) and (3.48)) can amount to a complete replacement of the previously ascribed state by a new one that contains no trace of the old ascription. This is prominently the case in the Stern-Gerlach experiment: For an atom coming out of the upper (or lower) channel of such a device oriented along the spatial direction $\vec{n}$, the adequate spin state to ascribe is $|\uparrow, \vec{n}\rangle$ (or $|\downarrow, \vec{n}\rangle$), irrespective of which state, if any, one had previously ascribed at the input side of the device. This shows the limits of thinking of state ascriptions in terms of “information about the system”, and the power of thinking in terms of an agent's interaction with the system. In stark contrast to the classical case, an agent's “information about the system” does not always accumulate smoothly.

One can think of an experiment as asking a specific question about a system and recording one of several possible answers as the outcome. This attitude is valid both for classical physics and for QM. But what we just said highlights an important difference. In classical physics, the experimental questions can be understood as being about something that is *already there*. Measurement *reveals* the property one is interested in. True, measurement often disturbs a system—if you measure the temperature of a hot fluid by immersing a cooler thermometer, the thermometer has to thermalise, and this will cool the liquid a bit. But one can take care of these disturbances when analysing the result, and an ideal measurement will cause only negligible disturbance. On that classical understanding, information about different properties of a system will simply accumulate. Also, there is no theoretical limit to the precision with which one can determine values of different properties together. For example, for a classical spinning body, information about the angular momentum with respect to any chosen axis can be consistently integrated into one coherent picture.

The upshot of the study of sequential Stern-Gerlach experiments, however, is that information about the spin alignment along different directions does *not* accumulate. Classically, one would assume that an atom leaving the oven has its spin aligned along some direction $\vec{n}$, and that a Stern-Gerlach experiment somehow takes recourse to the projection of that direction along, e.g., the x - or the z -axis in determining the outcome. But in the actual experiment, the alignment seems to be completely *flipped* one way or the other through the measurement interaction—after passing the second apparatus pictured in Fig. 3.2b, the description of the atomic state valid at its input, $|\uparrow, z\rangle$, is *replaced* by one of $|\uparrow, x\rangle$ or $|\downarrow, x\rangle$, wiping out the

initial alignment along z . It seems that the questions “how is the spin of the atom aligned along z ” and “how is the spin of the atom aligned along x ” cannot both receive an answer with respect to the same system at the same time. In orienting the apparatus one way or the other, one chooses to pose one of these questions and not the other, and the one asks receives an answer, while the question that was not asked does not even receive a hypothetical, counterfactual answer. Given the atomic state $|\uparrow, z\rangle$, one knows all about the alignment of the atom along the z -direction (one is guaranteed to find parallel alignment, never antiparallel alignment if one checks via a subsequent z -oriented apparatus), but one knows *nothing* about the alignment along the x -direction, and accordingly, the outcome of a measurement along the x -direction is completely random. In fact, a set-up of roughly this form is used in the best technological means for producing randomness; see our discussion in Sect. 3.5.3 below.

To check for the randomness of the outcome of an x -measurement for a system prepared in a z -eigenstate, we can use the Born rule for kets (Eq. (3.43)) and the explicit ket representation of Eq. (3.71):

$$\Pr(\uparrow, x) = |\langle \uparrow, x | \uparrow, z \rangle|^2 = \left| \left(\frac{1}{\sqrt{2}} \begin{pmatrix} 1 \\ 1 \end{pmatrix}, \begin{pmatrix} 1 \\ 0 \end{pmatrix} \right) \right|^2 = \frac{1}{2} = \Pr(\downarrow, x). \quad (3.91)$$

It was in order to describe such extreme cases of incompatibility of observables, which imply total ignorance regarding one of them, that Bohr coined the term “complementarity”—another much discussed and controversial notion in QM. At its most basic, complementarity implies that there are meaningful questions about a system that cannot be fully answered together, even though each of the questions separately *can* be answered. As we have seen, spin projections along perpendicular axes (such as x and z) are perfect examples of complementarity. Famously, position and momentum also constitute a complementary pair in QM, while their classical treatment as conjugate variables contains no hint of incompatibility. The existence of complementary observables is a specifically quantum phenomenon. What lies behind it?

Historically, in the work of Heisenberg and Bohr and others starting in the 1920s, the focus was on arguments for why position and momentum should be complementary. One may argue, for example, that the exact determination of the position of a particle presupposes an ideally rigid reference frame, whereas part of a measurement system must be non-rigid and movable in order to determine momentum.⁷² But such classically inspired arguments can be misleading, and anyway they do not carry over to other pairs of complementary observables such as spin projections. Complementarity is a pervasive phenomenon in QM. In our view, the most useful perspective on this issue is to acknowledge that complementarity is already deeply ingrained in the mathematical structure of QM, as it maps onto a basic feature of the operator algebra. Mathematically, complementarity is a case of

⁷² See Hilgevoord and Uffink (2024) for a useful overview of the history of these considerations.

the non-commutativity of the respective operators (observables), which we already pointed to when defining the commutator of two operators above (Eq. (3.29)). Agents who use QM as a tool are well advised to accept complementarity as a basic feature of the tool they are using.⁷³

It may be instructive here to illustrate the connection between the complementarity of different spin components and its mapping onto the corresponding operator algebra. The mathematical expression (Schwinger, 2001) of the complementarity of the different spin components is given by a set of relations as in Eq. (3.76):

$$\begin{aligned} |\langle j, x|i, z\rangle|^2 &= \frac{1}{2}, \\ |\langle j, y|i, z\rangle|^2 &= \frac{1}{2}, \\ |\langle j, x|i, y\rangle|^2 &= \frac{1}{2}, \end{aligned} \tag{3.92}$$

for any choice of $i, j \in \{\uparrow, \downarrow\}$. These relations express the basic conditional probabilities that are observed in experiments with Stern Gerlach devices as described in Fig. 3.2. The first of these relations, for example, states that for a system whose spin is prepared along the z -direction the outcome of a subsequent measurement along the x direction is unpredictable (i.e., both outcomes have the same probability of 1/2). At the same time, the outcome of a measurement along the z direction is predictable with certainty (see Eq. (3.75)).

The relations in Eq. (3.92) together imply algebraic relations for the corresponding observables $\hat{S}_x$, $\hat{S}_y$, and $\hat{S}_z$. For example,

$$[\hat{S}_z, \hat{S}_x] = \hat{S}_z \hat{S}_x - \hat{S}_x \hat{S}_z = i\hbar \hat{S}_y, \tag{3.93}$$

and similar relations hold for cyclic permutations of the indices z, x, y . From the relations Eq. (3.92) and the choice Eq. (3.70), one can easily obtain the matrix representations of the spin observables (Eq. (3.72)) and their corresponding eigenstates (Eqs. (3.70) and (3.71)), from which Eq. (3.93) follows straightforwardly. The algebraic relations in Eq. (3.93) are thus a direct consequence of the complementarity of the different spin components as expressed in Eq. (3.92). Those relations enter in Eq. (3.80) and thereby determine the dynamics and time evolution of the atomic spin in a magnetic field. They also tell us something about the degree of

⁷³ An agency perspective can even offer useful, if limited, classical analogies of situations in which perfectly meaningful questions cannot be answered together. Suppose you are at home with no specific obligations, and you deliberate whether to cook or to read, settling for reading. Then while there is a definite answer to what you are reading, there is no definitive answer to how much olive oil you would have put into the pasta sauce that, in fact, you did not make, nor indeed on whether it would have been pasta you would have cooked. We will later see how considerations of the openness of the future, which agents are facing all the time, are relevant for elucidating other conceptual problems of QM as well.

incompatibility of definite spin components along any two directions, even when they are not necessarily orthogonal.⁷⁴ The implications of the complementarity relations are thus far reaching, and there are approaches to quantum mechanics (Schwinger, 2001; Englert, 2006a) that develop the entire mathematical structure of quantum mechanics largely from complementarity relations as basic empirical facts established by the Stern-Gerlach experiment.

From our perspective of treating QM as a tool for agents to get around, it is especially helpful to see that one can quantify the degree of mutual incompatibility of observables and the corresponding lack of possible co-determination of values quite generally, not restricted to observables that are fully complementary. This is the subject of the various so-called (Heisenberg) uncertainty relations (see Atkinson & Peijnenburg, 2022, for a recent overview). The earliest result along those lines, derived by Heisenberg (1927), says that for any particle, the product of the respective uncertainties or spreads in the determination of position and momentum must be at least of the order of Planck's constant, h .

With a view to a general result, for any observable (Hermitian operator) $\hat{A}$ and quantum state ρ we can define the quantity $\Delta\hat{A}$ to be

$$\Delta\hat{A} = \hat{A} - \langle\hat{A}\rangle\hat{1}, \quad (3.94)$$

with $\langle\hat{A}\rangle = \text{tr}(\hat{A}\rho)$. Then the *dispersion* (or *variance*) of $\hat{A}$ (that is, the variance of the probability distribution $\text{Pr}(a_i) = \text{tr}(\hat{P}_{a_i}\rho)$ associated with $\hat{A} = \sum_i a_i \hat{P}_{a_i}$) is the expectation value of the square of $\Delta\hat{A}$, which can be written as

$$\langle(\Delta\hat{A})^2\rangle = \langle(\hat{A}^2 - 2\hat{A}\langle\hat{A}\rangle + \langle\hat{A}\rangle^2)\rangle = \langle\hat{A}^2\rangle - \langle\hat{A}\rangle^2. \quad (3.95)$$

For any two observables $\hat{A}$ and $\hat{B}$, one can then derive an uncertainty relation that provides a lower bound on the product of the respective dispersions:

$$\langle(\Delta\hat{A})^2\rangle\langle(\Delta\hat{B})^2\rangle \geq \frac{1}{4} |[\hat{A}, \hat{B}]|^2. \quad (3.96)$$

Accordingly, for commuting observables ($[\hat{A}, \hat{B}] = 0$), both observables can in principle have dispersion-free values, whereas the inequality constrains the joint determination of position and momentum, or of spin projections $\hat{S}_x$ and $\hat{S}_z$. For the latter case, we obtain

$$\langle(\Delta\hat{S}_z)^2\rangle\langle(\Delta\hat{S}_x)^2\rangle \geq \frac{\hbar^2}{4} |\langle\hat{S}_y\rangle|^2. \quad (3.97)$$

⁷⁴ Through the linearity of the commutator one can easily show that, for any two directions given by unit vectors $\vec{a}$ and $\vec{b}$, the spin components along these directions, i.e., $\vec{a} \cdot \hat{\vec{S}}$ and $\vec{b} \cdot \hat{\vec{S}}$, satisfy the commutator relation $[\vec{a} \cdot \hat{\vec{S}}, \vec{b} \cdot \hat{\vec{S}}] = i\hbar(\vec{a} \times \vec{b}) \cdot \hat{\vec{S}}$.

If we consider an atom whose spin is prepared in the state $\rho = |\uparrow, y\rangle\langle\uparrow, y| = (\hat{1} + \sigma_y)/2$, then $\langle\hat{S}_y\rangle = \hbar/2$ which constrains the variances of the spin components in the x and z direction according to

$$\langle(\Delta\hat{S}_z)^2\rangle\langle(\Delta\hat{S}_x)^2\rangle \geq \frac{\hbar^4}{16}. \quad (3.98)$$

Indeed, for this case, we can easily compute $\langle(\Delta\hat{S}_z)^2\rangle = \langle\hat{S}_z^2\rangle = \hbar^2/4$, since $\hat{S}_z^2 = (\hbar^2/4)\hat{1}$ and $\langle\hat{S}_z\rangle = \text{tr}(\rho\hat{S}_z) = \langle\uparrow, y|\hat{S}_z|\uparrow, y\rangle = 0$. Similarly, $\langle(\Delta\hat{S}_x)^2\rangle = \hbar^2/4$. Thus the inequality Eq. (3.97) becomes satisfied by an equality.

Similarly, for position and momentum (say, in the x -direction), the commutator is $[\hat{X}, \hat{P}_x] = i\hbar\hat{1}$, from which one obtains the uncertainty relation

$$\langle(\Delta\hat{X})^2\rangle\langle(\Delta\hat{P}_x)^2\rangle \geq \frac{\hbar^2}{4} \quad (3.99)$$

that we already met in our discussion of the spreading of the wave package in Sect. 3.3.2.2.

3.4 The Deterministic Itch

As we just saw, one consequence of complementarity is base-level randomness, even under maximal information about a system: If an agent is justified in ascribing to a system an eigenstate of $\hat{S}_z$, her prediction for a measurement of $\hat{S}_x$ must be completely random, and this is borne out experimentally, not just with respect to the frequency of the two possible results of an $\hat{S}_x$ -measurement, but also with respect to all sorts of correlations between consecutive results. This base-level randomness is not chaos—on the contrary, the individually random outcomes happen with a perfectly reproducible, i.i.d. 50:50 statistics. And this fact of stable, reproducible statistics holds not just in the case of the complementary observables $\hat{S}_x$ and $\hat{S}_z$, but quite generally, as a consequence of the Born rule. For example, in the set-up of Fig. 3.2, by turning the second apparatus away from the alignment of the first by an angle φ , one can tune the statistics to show any possible ratio between up- vs. down-outcomes, quantified as $\text{Pr}(\uparrow, \varphi) = \cos^2(\varphi/2)$ (see Eq. (3.3)). Some interpreters of QM find the combination of base-level randomness with law-like statistics worrying. We describe the dialectics of these worries here under the heading of “the deterministic itch”.

3.4.1 How to Deal with Quantum Randomness?

There are different reactions to the fact that many experiments, including the Stern-Gerlach experiment, exhibit basic randomness with reliably reproducible statistics.

One reaction is to accept this as an insight into what nature is like, and to welcome it as a novel resource, e.g., for building new technologies. (We will describe some relevant technologies in Sect. 3.5 below.) In the development of science, we have come to adopt quite a number of insights into what nature is like. For example, we have come to accept that at the level of chemistry, matter consists of chemically indivisible atoms, and that these atoms can be sorted into a small number of basic kinds whose members are *exactly* alike—quite unlike the members of biological kinds.⁷⁵ We have also come to accept that these different chemical atoms are all made up of only three distinct elementary particles—protons, neutrons, and electrons, which in turn are also all exactly alike. The fact that these minute entities are intrinsically indistinguishable shows up in their statistics, giving rise to strange phenomena like Bose-Einstein condensation. With respect to fundamental quantum randomness, humble acceptance seems to be the most useful attitude from an agency perspective, and we will go on to defend it. But there is also a rationally defensible response that goes in a different direction, viz., to question whether the randomness that we find is really fundamental, or whether it can perhaps be explained away at some deeper level. Such an explanation would show that behind the random variations in the individual outcomes of quantum experiments at the surface level, there is a hidden cause that influences the outcomes, perhaps to the extent that there is no randomness any more when things are viewed at that deeper level.

The urge to find such a deeper level is what we here call “the deterministic itch”. In QM, it is also known as the quest for finding so-called hidden variables. This is another highly debated area of quantum foundations, and we will not be able to give a comprehensive overview here, much less satisfy the proponents of all camps.⁷⁶ We want to start, however, with something that should be uncontroversial: The quest for deeper explanations of surface phenomena has been a driving force behind many important advances in science, and it is a well-tested heuristic for anyone who wants to understand things. The basic idea is that a stochastic surface phenomenon—whether with or without stable probabilities—can often be viewed as the result of a coarse-graining of an underlying system behaviour that may well be deterministic. Via the coarse-graining, information is lost. And that loss of information might be all there is to the surface-level stochasticity in the end. The way to find out about such an underlying layer is often via experiment, paired with careful observation: by varying the experimental conditions, one will hopefully also be systematically varying those factors that cause the surface-level variations, so that one can then advance to a deeper, more fine-grained understanding of the phenomena based on these factors. This method is in fact familiar from everyday examples—if you develop a skin rash on some days but not on others, you can vary your diet to find

⁷⁵ Apparently, ancient theories of atomism instead could allow for manifold individual differences among the atoms. Thus, for Democritus, atoms as indivisibles did not need to fall into a few neatly circumscribed kinds. See Berryman (2022, §2.1) for an overview.

⁷⁶ We find that Myrvold (2022) provides a balanced and highly informative overview of philosophical issues in QM, including but not limited to hidden variables.

out which food items you might be allergic to, so that you can hope to live without a skin rash in the future. In the history of science and medicine, Semmelweis's experiments that led to the introduction of hand-washing to reduce infections stand out as an example in which insight into the existence of a hidden factor predated a theoretical explanation of that factor by many years.⁷⁷

If one pushes the idea of hidden variables with respect to the silver atoms in the Stern-Gerlach experiment, one may come up with the idea that underneath the state ascriptions that QM offers via Hilbert space ket vectors such as $|\uparrow, z\rangle$, there should be a more complete description in terms of additional parameters λ not treated by QM, such that the true state of an atom can be written as $|\uparrow, z; \lambda\rangle$ (which cannot be a vector in two-dimensional Hilbert space, but must belong to some larger abstract state space). Given such a hypothetical complete, deeper-level state, we might be in a position to reinstate the fundamental classical relation between states and properties, doing away with the merely probabilistic Born rule: any possible property of the system, such as its spin-projection along the x -axis, would then be a function of that state $|\uparrow, z; \lambda\rangle$. In this way, the deterministic itch in QM urges one to search for such complementing parameters, or hidden variables, λ . Can they be found?

In principle this appears to be an empirical question, and so, maybe only time will tell. In the history of the sciences, theories come and go, and a well-confirmed attitude towards these changes is expressed by Laudan's (1981) pessimistic meta-induction: especially with respect to the underlying ontology, change is often radical and impossible to predict. But in the case at issue, QM has at least withstood all attempts at replacement by a purported completion since its inception in the 1920s, for almost 100 years.⁷⁸ Of course, such historical considerations get one only so far. Are there other ways to deal with the deterministic itch in QM?

The history of relevant attempts is interesting, but also tangled, so that even a mere presentation of that history may already be controversial. To mention a

⁷⁷ The explanatory theory was Pasteur's germ-account of infection, providing posthumous justification of Semmelweis's prescriptions, which had been violently rejected by his contemporaries (see Obenchain, 2016). The case of Semmelweis also demonstrates that the underlying level of understanding need not be deterministic; models for the transmission of infectious diseases are stochastic, but still explanatory. See also Hacking (1990) for an account of how, in the later nineteenth century, statistical thinking became acknowledged as having real explanatory power, first in the social sciences and then also in the natural sciences.

⁷⁸ As this remark will probably be met with resistance by adherents of Bohmian mechanics, we offer a few remarks here. Bohmian mechanics is often described as a deterministic alternative to QM that makes the same empirical predictions. The fine print is that this holds only for non-relativistic QM in which particle creation and annihilation is absent. Once these effects—which are central, e.g., to the fusion processes in the sun—are taken into account, determinism cannot be upheld; see Footnote 40 to Chap. 2. The discussion then reverts to the question of whether an alternative to QM such as Bohmianism is superior because it offers “an ontology”. In our view, this line of approach threatens to become an exchange of strongly held personal predilections. For what it's worth, we already put our own predilections on the table in Chap. 2: An ontology that cannot incorporate the reality of agents as ontologically emergent entities lacks the resources to describe experiments as experiments, and polemically, one could therefore call such an ontology antiscientific.

few main points: In the early 1930s, Von Neumann (1932) published an important book on the mathematics of QM in which he claimed to provide a proof of the impossibility of hidden variables. The assumptions behind that proof are nowadays generally criticised as unwarranted and too strong. In 1935, Einstein, Podolsky and Rosen (“EPR”) published their landmark paper “Can quantum-mechanical description of physical reality be considered complete?” (Einstein et al., 1935), which addresses the hidden variable question from the point of view of quantum correlations. Such correlations have been the focus for much of the debate since then. EPR argue that predicted correlations in multipartite systems, which are based on so-called entanglement of the sub-systems, call out for an explanation that can only be given by a hypothetical “completion” of QM, which amounts to adding hidden variables to QM. The assumptions behind EPR’s argument, especially their criterion of reality that licenses ontological consequences from probability-1 statements, are still debated. The next major step in the discussion was a theorem on the structure of the QM state space by Gleason (1957), which was the basis for what is known as the (Bell-)Kochen-Specker-theorem (Bell, 1966; Kochen & Specker, 1967). That theorem states, roughly, that for a Hilbert space of dimension three or larger, one cannot construct a theory with deterministic predictions by fine-graining the quantum state vectors in any consistent way.⁷⁹ This means that no hidden variables can be admitted without deviating substantially from the mathematics of standard QM, but it still leaves open the option to implement hidden variables via a radical change away from that accepted mathematical structure. The most decisive advance then came with John Bell. He showed how hidden variables of certain types constrain the range of possible empirical phenomena, ruling out correlations above a certain strength. These results are often collectively called “Bell’s theorem”, codified via a series of so-called Bell inequalities that express limits for the strength of correlations supported by hidden variable models of certain types.

The advantage of Bell’s approach is that it does not require a purported hidden-variable “completion” of QM to fine-grain the actual mathematical structure of that (historically contingent) theory. It thereby enables an inference from observed correlations directly to the structure of all possibly underlying theories.⁸⁰ Hidden variables are not generally impossible—Bell (1966, §2) in fact starts by explicitly constructing a hidden variable theory for the results of spin measurements on a spin-1/2 particle, such as a silver atom in the Stern-Gerlach experiment. That is, he shows that one could, if one were so inclined, give in to the deterministic itch with respect

⁷⁹ For Hilbert spaces of dimension four and larger, 18 observables are sufficient to establish a contradiction from the assumption of deterministic predictions; see Cabello et al. (1996). This is indeed the minimal number for dimension four (Pavičić et al., 2005). In three dimensions, Peres (1991) proves the theorem using 33 observables. The original proofs from the 1960s were more complex.

⁸⁰ Remaining controversies either target loopholes in experiments, e.g., due to less-than-ideal detectors, or criticise the validity of the assumptions about the class of hidden variables that is taken into consideration. Explicit discussion of the agency aspect of experimentation only rarely play a role. See Müller and Placek (2023) for some pertinent discussion.

to that experiment, and posit alternative states for the atoms. More importantly, however, Bell shows that this is a peculiarity of the case: the Hilbert space of dimension two, which describes a spin-1/2 particle, is really an outlier. (As we said above, the Bell-Kochen-Specker-theorem is valid only for Hilbert space dimensions three or larger.) For systems of two entangled spin-1/2 particles (with Hilbert space dimension four), Bell (1964) proves that a broad class of hidden variable models is ruled out by the correlations predicted by QM.⁸¹ This can be understood as the basis for putting the question of the existence of hidden variables to an empirical test. Who is right? Experiments were infeasible in the 1960, but started in the 1970s, with a landmark experiment by Aspect et al. (1982) among those strongly confirming the predictions of QM.⁸² These experiments, so-called Bell tests, have meanwhile evolved considerably. On the one hand, various loopholes that were present in earlier experiments have been closed.⁸³ On the other hand, techniques for creating and manipulating entangled quantum systems consisting, e.g., of photons or ions, have become widely available. Such techniques form the basis of a number of emerging technologies, for example for quantum communication, cryptography, and quantum computation. We will describe some of these technologies in Sect. 3.5.

Before we go on to discuss Bell's results in Sect. 3.4.3, we address some common conceptions of what makes a good theory, as such conceptions provide the motivation for assumptions that are made in the proofs of the various Bell theorems.

3.4.2 *Images of a Good Theory*

A theory is, first and foremost, a tool for agents to get around in the uncertain world—to predict future happenings, to enhance control over things of interest, or to build technological artefacts. When theories allow us to predict the future, the most general, precise question they can answer is this: Given what we know about the system we are interacting with at the present time, what can happen at some given later time, and what are the odds? An exhaustive answer to that question would consist of a list of possibilities for these future happenings with respective probabilities attached. QM gives us exactly that, via the Born rule. What more could one want?

⁸¹ The publication dates of the two relevant early publications by Bell are confusing. In fact, Bell (1966) was written *before* Bell (1964); the paper published in 1966 contains a cliff-hanger remark (Footnote 19) that announces the result published in 1964 (giving 1965 as publication year). The reason for this infelicitous ordering was an editorial mishap that led to delayed publication of the earlier paper.

⁸² Aspect's, Clauser's, and Zeilinger's early works on the theoretical and experimental vindication of the predictions of QM against hidden variables were awarded with the Nobel prize for physics in 2022.

⁸³ For relevant recent results, see Hensen et al. (2015), Giustina et al. (2015), Shalm et al. (2015), and Abellán et al. (2018).

When We Want More than a List of Possibilities More *is* often wanted, and often, but not always, for good reasons. Possibility contrasts with actuality. If one knows a range of possibilities, isn't that just an expression of uncertainty about what is actually so? And couldn't that uncertainty, at least in principle, be removed? This line of reasoning may sound convincing, but it betrays a conflation of two distinct types of uncertainty. These two types of uncertainty are, furthermore, aligned with two different types of contrast between possibility and actuality. A common predicament illustrates this contrast. Suppose you are looking for your keys. The keys are actually somewhere, but due to lack of knowledge, you may be confronted with a range of possibilities for where they might be. Their actual place is just one among these possibilities. In such a case, more information will eventually help to limit the set of possibilities to one, centring on actuality. The initial possibilities are merely epistemic *alternatives to* that actuality. But now consider, still searching for your keys, the question of when you will find them. There is uncertainty about that, too, but it is not as if there was a settled actuality that you just had to discover. Actually, you have to *do* something to find the keys, and possibly, you may find them in the next minute, in five minutes, or much later (possibly even never, if you abandon your search). These possibilities are not *alternatives to* actuality, but *alternatives for* the future, which is actually, as of the time of your search, still open.⁸⁴

Our agential uncertainty is often uncertainty about the really open future rather than uncertainty merely due to lack of knowledge. And before a really open issue is settled and an outcome is actualised, there is simply nothing there to be known. In such a case, the best even a perfect theory can do is to offer a list of possibilities with attached probabilities. And that is exactly the situation that we have in the case of quantum mechanics. The probabilistic nature of that theory is basic and tied to the openness of the future. As agents, we are dealing with such openness all the time, for example because other agents really settle what they do, which cannot be “the inevitable consequences of what has gone before” (Steward, 2012, 72). From this perspective, then, QM in its standard form is a model of a perfectly good theory. But the idea that the future is truly open is contested despite its agential credentials. There are at least two images of what a good theory should be that operate with the deterministic idea of one actual, unique future, thus marking any uncertainty as an epistemic defect: the clockwork universe and the bitmap analogy.⁸⁵

⁸⁴ See Rumberg (2016) for an explication of the contrast, and see Belnap et al. (2022, Ch. 1) for some further examples. The contrast between possibility and actuality can already be found in Aristotle (see *Metaphysics* 9, 1048^b4–6).

⁸⁵ In philosophy, an additional reason for dismissing the open future is the almost universal adoption of an analysis of possibilities in terms of possible worlds—a philosopher's paradise, as it has been called (Lewis, 1986a, Ch. 1). On that analysis, each possible world comes with a unique temporal course of development. Provided one has properly located oneself within the set of worlds, the future is therefore fixed. Any uncertainty is, on that view, ultimately epistemic, viz., a lack of proper self-location.

3.4.2.1 The Image of the Clockwork Universe

With respect to a few select phenomena like the motion of the planets, there are theories that successfully make unique predictions; Newton's gravitational theory is a prime example.⁸⁶ This has had a strong and lasting influence on our image of what a good theory should be. Besides the spectacular empirical success of the Newtonian theory, the specific subject matter of celestial mechanics surely also played a role in establishing it as a paradigm: The motions of the planets and other heavenly phenomena have for a long time been a source of inspiration and awe for humans, and knowing about them has often been a symbol of power. Calendars are important not just for agriculture, but also for anniversaries and religious festivities, and knowing the positions of sun, moon, planets, and stars is important for navigation, especially at sea. Furthermore, beliefs about the influence of celestial events on human affairs were and still are widespread. Many ancient cultures already had astonishingly precise knowledge of important aspects of astronomy. Crucial advances in observation and calculation were made in the sixteenth century. And with the advent of the telescope in the seventeenth century, new theories describing the apparent motion of the planets and the fixed stars could be put to empirical test. For example, Kepler in 1630 successfully predicted the 1631 transit of Mercury across the sun, which was indeed observed by Gassendi (Van Helden, 1976). Kepler's prediction came with its own uncertainties—he had given a range of days for the event, reflecting doubts about the numerics and the observations underlying his calculations. But subsequent refinements led to reliable, tabulated predictions: not a list of possibilities with probabilities attached, but one firm value for each future time. The resulting image of the universe as clockwork, as embodied by orreries (illustrative mechanical models, some actually driven by clockwork), has stuck. It continues to inform our idea of what successful theories must be like.

It is not so clear whether this is always a good thing. Anscombe, with whom we tend to agree, holds that “[t]he high success of Newton's astronomy was in one way an intellectual disaster: it produced an illusion from which we tend still to suffer”. The illusion is that any good theory must single out one unique future development for any system to which it applies, thus providing predictions with certainty. But as Anscombe (1971) notes, “it was mere obligingness on the part of the solar system, by having had so peaceful a history in recorded time, to provide such a model”. Other planetary systems, e.g., ones in which planets or moons explode or break apart, or in which the system's inherently chaotic dynamics leads to a loss of predictability over time scales relevant to observation, would probably not have led to the image of a clockwork universe.⁸⁷

⁸⁶ Some caveats apply; see Footnote 88 below for the fine print.

⁸⁷ Hayes (2007) estimates that the Lyapunov times for the motions of the outer planets of the solar system, which are roughly the times at which non-periodic effects show up, are in the order of a few million to a few hundred million years. By way of contrast, for the earth's atmosphere, that time is in the order of a few weeks. It is the difference in time scales rather than a fundamental

Apart from that contingency of origin, there is a further issue that concerns the scope of the clockwork image: A global theory that accords with that ideal could no longer be a tool for human agents facing worldly uncertainty. If it could be a tool at all, then it would have to be one for a supreme intelligence, like Laplace's famous demon, who knows everything about "the movements of the greatest bodies of the universe and those of the tiniest atom" by computing these movements from initial data, so that "the future just like the past would be present before its eyes" (Laplace, 1820, 4). The associated subtext is that the world is obviously deterministic, and so anything less than perfect prediction and retrodiction must be due to shortcomings of us humans in comparison with Laplace's demon: either the theory we use is wrong, or we do not know enough. Absent such deplorable limitations, there is no place for probabilities in a theory.⁸⁸

Even firm supporters of this view of course concede that often we cannot know enough, or that even if we could, we could not handle the immense amount of information that fully describes a system. Thus, probabilities are admitted in statistical theories like statistical mechanics, in which systems are described by a few coarse-grained variables that can be realised in many different ways. But the following is accepted as an item of faith: If a description of a system is maximally precise, and the theory is good, then the predictions of the theory will have the form of a single determinate value for each future time, rather than a list of possible values with associated probabilities. And this image is what clashes with QM—because the Born rule of QM "only" allows for probabilistic predictions, even given a maximally precise, pure quantum state. As we will discuss in Sect. 3.4.3, hidden variables are meant to remedy this alleged defect of QM by implanting something like an additional clockwork mechanism operative beneath the quantum state.

Deterministic dynamics need not lead to system behaviour that is regular like clockwork. In fact, in a sense that can be made quantitatively precise, deterministic systems can show behaviour that is as chaotic as one pleases Werndl (2011). The phenomenon of deterministic chaos is well studied in the theory of complex dynamical systems, and there are many examples of structurally simple, deterministic systems whose behaviour looks surprising. Cellular automata are a case in point. Their study has led to a further strong image of what good theories must be like.

difference of the dynamics that makes the weather system a proverbial example of uncertainty and the solar system a paradigm of stability.

⁸⁸ It is somewhat ironic that the metaphysical dream of Laplace's vision is not even borne out by the main contender for being a good theory at the time, Newtonian mechanics. For example, that theory does not make unique predictions for collisions of more than two particles and allows for infinite "space-invader" scenarios (Earman, 1986). Furthermore, there are even indeterministic Newtonian systems that do not rely on collisions or infinite arrangements (Xia, 1992).

3.4.2.2 The Bitmap Analogy

The motions of the heavens are regular, and it may seem reasonable to demand a simple deterministic theory for them. But the extrapolation of that ideal to all that happens anywhere in the world, which is implied by Laplace's vision, may still seem like a long shot not just because we could not handle it, but also given the general disorder of things here on earth. At this point, another analogy often replaces the picture of the universe as clockwork, one based on Conway's famous "Game of life". That algorithm, a cellular automaton, can generate surprisingly complex phenomena via a single, very simple and deterministic rule that operates on a two-dimensional grid of individual cells. At each time step, each cell can be on or off ("alive" or "dead"), and the total system state, a bitmap, determines the state at the next step, deriving a cell's next state from the states of its immediate neighbours. The analogy is that any complex system's state must be something like such a black-and-white bitmap: a mosaic of smallest, independently variable, simple parts that interact locally.⁸⁹

For such a bitmap, to have full information means to know of each pixel whether it is black or white—basically, to have a copy of the full system. That is maximally specific information, and when one has that, there is nothing more to be known: anything one can ask about the bitmap, including any purported higher-level features, must supervene on the state of the bitmap, which is philosophical jargon for saying that there can be no higher-level difference without a lower-level difference.⁹⁰ When one asks questions about the system that are not about the state of a single pixel of the bitmap, these are about aggregate features that must translate down to the lowest level of the pixels somehow. For example, one can ask what the parity of the map is (e.g., taking white = 0 = "off", black = 1 = "on"), and one computes the answer by summing over all pixels and checking whether the result is even or odd. Or one can display the bitmap as a two-dimensional picture and ask whether it anywhere shows a recognisable mathematical equation, like " $7+5 = 12$ ". In the latter case it will not be immediately obvious how to resolve this question in terms of the individual pixels, but it is clear that any solution will, in the end, have to address the values of these pixels, as that is all there is to the system. So, following

⁸⁹ The bitmap analogy plays a strong role in Dennett (2003) and other writings of that author. The superdeterministic "cellular automaton interpretation of quantum mechanics" by t'Hooft (2016) is driven by the same idea; see Footnote 148. And as we mentioned in Footnote 24 to Chap. 2, the popular image of the Star Trek transporter lives off the idea that a material object is essentially a 3D bitmap of independent atoms. The similar idea of a so-called *Humean mosaic* of independent "local matters of fact" also provides motivation for an influential research programme in philosophy, Lewis's so-called "Humean supervenience" (see Lewis, 1986b, ix).

⁹⁰ In the philosophical research programme mentioned in Footnote 89, the bitmap idea is generalised across the board. And indeed: if the bitmap—or the "Humean mosaic" of local facts—is all there is to any system (and, indeed, to the universe as a whole), then anything one can say about it—including assertions concerning dispositions and counterfactual "what would happen if" questions—"supervenes on that". This means that any difference in what one can say must come down to a difference in the mosaic, as there is nothing else that could make a difference.

this analogy, to have maximal information about a system will enable one to resolve all possible questions about the system's properties precisely and uniquely. And as the Game of life shows by example, even highly complex bitmap dynamics that include seemingly emergent features can be generated by a simple deterministic local rule. In this way, the bitmap analogy is used to suggest that determinism is compatible with the occurrence of all the messy complexity of the world we see around us.

For a bitmap, it is easy to understand what it is to have full information—basically, to have a copy of the system—, but also what it is to have less-than-full information, e.g., to have access to the bitmap at a reduced resolution. Having that may or may not be enough to resolve certain questions about the system (it will not be enough to decide the question of parity, but it may actually be helpful for solving the task of recognising a written equation). At any rate, the distinction between *information about the system* and the aggregate of the ultimate constituents of *the system itself* effectively collapses—the arrangement of constituents is all there is, and so its full description is basically the system twice over. This part of the analogy, as we will see, also plays an important role in motivating assumptions about hidden variables that lead to Bell's theorem.

The bitmap analogy and its associated notion of supervenience are widely accepted in philosophy, but in our view, two important points are lost. First, the bitmap analogy begs the question against ontological emergence, while we have no good scientific reasons for accepting it. If the world was really like a mosaic of independently variable basic constituents, some variant of the bitmap analogy would apply, and higher-level features would have to supervene on the bitmap's state. However, if there is in the world any instance of ontological emergence—for example, if the sun is real—, then the bitmap analogy must be misleading, and the associated image of Humean supervenience has to fail. Secondly, the bitmap analogy dismisses our leading idea that theories should be tools for agents to get around. True, an agent of sufficient capabilities could just work with a copy of the system she is interested in—a “digital twin” perhaps—to resolve all questions. This could have many advantages compared to working with the system itself—it could be cheaper, faster, more practical, safer, or ethically less problematic. And of course it is fine to use the image of a system as a bitmap where appropriate. For example, that image provides an excellent framework for describing what goes on in a digital computer's memory, which *is*, in effect, a bitmap, and whose dynamics is described by deterministic local rules. The Ising model of ferromagnetism is a further example of a helpful use of the bitmap analogy; artificial neural networks are another case in point. But for large and complex systems, the analogy saddles an agent with an impossible task and provides no real help for how to get around. For this reason, we are for example critical of approaches to climate modelling via a “digital twin earth” (Bauer et al., 2021). For an agent facing our messy world, it does not help to assume that one has access to a copy of the system one is dealing with. One *is* already dealing with a system that is too large and complex to handle; what would access to a copy of that system contribute?

In dealing with a real, complex system, agents need to create simplifying tools—models or representations that, by necessity, include effects of perspective and idealisation. Agents cannot deal with the world “from nowhere”. As we will discuss in Sect. 4.4, this means that agents need to model or simulate those aspects of the world that they cannot oversee themselves, taking into account their individual perspective rather than trying to work from an all-encompassing non-perspective. As we will see in discussing Bell’s results, disregarding that perspectival nature of modelling is one main source of philosophical confusion about quantum state ascriptions.

3.4.3 *Bell’s Theorem*

We have provided an overview of ideals of a theory because we believe that it forms a useful backdrop for a discussion of Bell’s theorem. Roughly, that theorem (or class of theorems) says that quantum systems cannot be understood as aggregates of independently variable parts undergoing deterministic local evolution, while common assumptions about hidden variables force one to try to do exactly that.

Bell on Hidden Variables We already agreed that it may be rational to look for deeper-level accounts of stochastic surface phenomena in the hope for more satisfactory explanations of what goes on in the world. Bell’s research was in part triggered by the fact that a deeper-level, hidden variable account of many stochastic quantum phenomena is actually available: Bohm (1952) had provided an updated version of a deterministic approach that had already been discussed in the early days of the development of QM, the so-called pilot wave theory of de Broglie. Bohm’s theory gets rid of the indeterminism of (non-relativistic) QM, but its explanatory value was and still is debated. Bell had studied Bohm’s theory and established that it provides a counterexample to a number of purported “no hidden variables”-proofs. Still, Bell found it “extraordinary” as a physical theory, and he knew that few physicists took it seriously. In fact, according to Bell, Bohm’s theory provided a completion of quantum mechanics, and therefore a resolution of the EPR paradox, “in the way which Einstein would have liked least” (Bell, 1966): It posits a causal mechanism that relies on a kind of instantaneous, non-local action-at-a-distance.⁹¹

Having convinced himself that hidden variable theories of this (unsatisfactory) kind are possible, Bell went on to ask whether “any hidden variable account of quantum mechanics *must* have this extraordinary character” (ibid.). His proof (Bell, 1964) provides a broadly positive answer: The introduction of hidden variables leads to predictions that deviate from those of QM unless the hidden variable theory violates a certain principle of locality. Bell captures this principle via a constraint on the probability distributions that govern the theory’s predictions about distant correlations.

⁹¹ This happens only behind the scenes at the level of hidden variables beyond experimenters’ control, not at the level of observable phenomena. Therefore, no empirical clash with relativity theory is implied.

In a large number of publications, the full set of assumptions behind Bell's early proofs and the further arguments he gave later on have been scrutinised, and others have contributed to the debate by devising proofs that make fewer or weaker assumptions. A useful, well-balanced overview of the relevant literature is given by Myrvold et al. (2024). We will not provide a historical overview or an in-depth discussion here, but limit ourselves to a brief presentation of a Bell-type argument before drawing our conclusions. We start with some preliminaries about the mathematical description of entangled composite systems.

Entanglement and the Bell Singlet State Bell's theorem relies on the quantum-mechanical phenomenon of entanglement, which, as we said, was also behind the original 1935 EPR argument (Einstein et al., 1935). Entanglement is a purely quantum phenomenon that is characteristic of correlated composite quantum systems (Schrödinger, 1935). Such systems can exhibit correlations for which there is no classical analogue. For Bell's theorem, it suffices to consider a two-partite system, and so we will focus on that case—even though entanglement is pervasive also in much larger systems.

Mathematically, in order to describe the joint state of a two-partite system composed of individual systems S_1 and S_2 , one has to use the tensor product of the Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$ that are appropriate, respectively, for system 1 and 2:

$$\mathcal{H}_{12} = \mathcal{H}_1 \otimes \mathcal{H}_2. \quad (3.100)$$

That larger Hilbert space, of dimension $d_{12} = d_1 d_2$, contains all linear combinations of so-called product states $|\psi_1\rangle \otimes |\psi_2\rangle$. Such linear combinations need not be product states themselves. If we consider two spin-1/2 particles, such as the silver atoms in the Stern-Gerlach-experiment, each described by a Hilbert space of dimension two, their tensor product has dimension four, and it contains both *product* states, such as

$$|\uparrow, z; \downarrow, x\rangle_{12} \equiv |\uparrow, z\rangle_1 \otimes |\downarrow, x\rangle_2, \quad (3.101)$$

and *non-product states*, such as the so-called *Bell singlet state*

$$|\Psi_S\rangle_{12} = \frac{1}{\sqrt{2}} (|\uparrow, z\rangle_1 \otimes |\downarrow, z\rangle_2 - |\downarrow, z\rangle_1 \otimes |\uparrow, z\rangle_2). \quad (3.102)$$

We can check that this state is rotationally symmetric, showing no preferred direction, so that, for any unit vector $\vec{n}$ we have (up to a global phase factor):

$$|\Psi_S\rangle_{12} = \frac{1}{\sqrt{2}} (|\uparrow, \vec{n}\rangle_1 \otimes |\downarrow, \vec{n}\rangle_2 - |\downarrow, \vec{n}\rangle_1 \otimes |\uparrow, \vec{n}\rangle_2). \quad (3.103)$$

For a proof, note that $|\uparrow, \vec{n}\rangle$, as a vector in two-dimensional Hilbert space, must be a linear combination of $|\uparrow, z\rangle$ and $|\downarrow, z\rangle$, which together form a basis. So, for some complex coefficients α and β , we have $|\uparrow, \vec{n}\rangle = \alpha|\uparrow, z\rangle + \beta|\downarrow, z\rangle$. Up to a

choice of phase, this implies $|\downarrow, \vec{n}\rangle = -\beta^*|\uparrow, z\rangle + \alpha^*|\downarrow, z\rangle$. Equation (3.103) then follows by substituting and collecting terms, noting that due to normalisation, we have $\alpha\alpha^* + \beta\beta^* = \|\uparrow, \vec{n}\|^2 = \|\downarrow, \vec{n}\|^2 = 1$.⁹²

It is also instructive to check that state indeed is not a product state, i.e., cannot be written as a tensor product. This can be most easily done by taking the *partial trace* over the state of one system. Partial trace is an operation that becomes relevant when we consider the parts of multi-partite systems, for example, when individual component systems are measured by different agents, as discussed below. The operation is also fundamental in quantum information applications of entangled states (see Sect. 3.4.4), and it plays an important role in the discussion of noise, e.g., in the theory of quantum error correction (Sect. 3.5.4).

Consider the general case of a two-partite system as introduced above, where $\{|a_j\rangle_1 | j = 1, \dots, d_1\}$ and $\{|b_k\rangle_2 | k = 1, \dots, d_2\}$ are sets of basis states for the Hilbert spaces $\mathcal{H}_1$ and $\mathcal{H}_2$, respectively. Thus the set of tensor-product states $\{|a_j\rangle_1 \otimes |b_k\rangle_2 | j = 1, \dots, d_1, k = 1, \dots, d_2\}$ provides a basis of the total Hilbert space $\mathcal{H}_1 \otimes \mathcal{H}_2$. Any state of the two-partite system can thus be written in the form

$$|\Psi\rangle_{12} = \sum_{j=1}^{d_1} \sum_{k=1}^{d_2} c_{jk} |a_j\rangle_1 \otimes |b_k\rangle_2, \quad (3.104)$$

of which Eq. (3.102) is an example. The generalization of Eq. (3.104) to an N -partite system is straightforward. If the system is in a product state, as in Eq. (3.101), then each subsystem can be separately assigned a pure state, namely the corresponding factor appearing in the tensor product.

Such an assignment of a quantum state to a part of a system can be generalized to multi-partite systems in an arbitrary state via the formal operation of the partial trace, which is defined on density operators describing the total system. If one considers a density operator ρ_{12} on the Hilbert space $\mathcal{H}_{12}$, then the partial trace over system 2, for example, gives the *reduced state* $\rho_1 = \text{tr}_2(\rho_{12})$ of system 1. The partial trace is defined as follows:⁹³

$$\text{tr}_2(\rho_{12}) = \sum_{k=1}^{d_2} {}_2\langle b_k | \rho_{12} | b_k \rangle_2. \quad (3.105)$$

For the density operator $\rho_{12} = |\Psi\rangle_{12}\langle\Psi|$ corresponding to the pure state $|\Psi\rangle_{12} = |\uparrow, z; \downarrow, x\rangle_{12} = |\uparrow, z\rangle_1 \otimes |\downarrow, x\rangle_2$ of Eq. (3.101) we obtain

⁹² For an alternative proof, one can show that the state in Eq. (3.102) is not changed when rotating both spins 1 and 2 around the same axis and by the same angle. Physically, such rotations would be generated, e.g., by applying a magnetic field as discussed in the context of Eqs. (3.80) and (3.81).

⁹³ Recall that $\{|b_k\rangle_2 | k = 1, \dots, d_2\}$ is an arbitrary basis for $\mathcal{H}_2$. We write “ ${}_2\langle b_k |$ ” for the bra corresponding to the ket $|b_k\rangle_2$, with the subscript indicating which component of the multipartite system is at issue.

$$\begin{aligned}
\rho_1 &= \text{tr}_2(|\uparrow, z; \downarrow, x\rangle_{12} \langle \uparrow, z; \downarrow, x|) \\
&= \text{tr}_2(|\uparrow, z\rangle_1 \langle \uparrow, z| \otimes |\downarrow, x\rangle_2 \langle \downarrow, x|) \\
&= |\uparrow, z\rangle_1 \langle \uparrow, z| \text{tr}(|\downarrow, x\rangle_2 \langle \downarrow, x|) \\
&= |\uparrow, z\rangle_1 \langle \uparrow, z| \langle \downarrow, x | \downarrow, x \rangle_2 \\
&= |\uparrow, z\rangle_1 \langle \uparrow, z|,
\end{aligned} \tag{3.106}$$

which is just the (projector onto the) pure state corresponding to the subsystem 1, as discussed above. For the entangled Bell singlet state of Eq. (3.102), $\rho_{12} = |\Psi_S\rangle_{12} \langle \Psi_S|$, the reduced state has the form

$$\begin{aligned}
\rho_1 &= \text{tr}_2(|\Psi_S\rangle_{12} \langle \Psi_S|) \\
&= \frac{1}{2} \text{tr}_2(|\uparrow, z\rangle_1 \langle \uparrow, z| \otimes |\downarrow, z\rangle_2 \langle \downarrow, z|) \\
&\quad - \frac{1}{2} \text{tr}_2(|\uparrow, z\rangle_1 \langle \downarrow, z| \otimes |\downarrow, z\rangle_2 \langle \uparrow, z|) \\
&\quad - \frac{1}{2} \text{tr}_2(|\downarrow, z\rangle_1 \langle \uparrow, z| \otimes |\uparrow, z\rangle_2 \langle \downarrow, z|) \\
&\quad + \frac{1}{2} \text{tr}_2(|\downarrow, z\rangle_1 \langle \downarrow, z| \otimes |\uparrow, z\rangle_2 \langle \uparrow, z|) \\
&= \frac{1}{2} (|\uparrow, z\rangle_1 \langle \uparrow, z| + |\downarrow, z\rangle_1 \langle \downarrow, z|) = \frac{1}{2} \hat{1}_1,
\end{aligned} \tag{3.107}$$

where we have used that, e.g., $\text{tr}_2(|\downarrow, z\rangle_2 \langle \uparrow, z|) = \langle \uparrow, z | \downarrow, z \rangle_2 = 0$. The last equality follows from Eq. (3.90). If the Bell singlet state were a product state, the reduced state of system 1 would be a pure state rather than a mixture.

Suppose now that we have a source of particle pairs that prepares them in the Bell singlet state, as depicted in Fig. 3.5, and that the individual particles fly apart, ending up, respectively, in the labs of Alice and Bob. It is interesting to see what happens when they perform measurements on their part of the Bell pair.

How Alice and Bob Update Their States In their respective labs, Alice and Bob are, as agents, free to choose their experimental conditions. We can assume that the

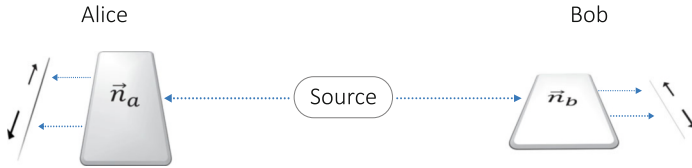


Fig. 3.5 Bell-type experiment with spin-entangled atoms in the Bell singlet state, to be measured by Alice and Bob via Stern-Gerlach devices in their respective labs

experiment concerns a pair of spin-entangled atoms, labelled a and b , respectively, flying in opposite directions along the y -axis, and that Alice and Bob each have a Stern-Gerlach device that they can rotate to any position in the xz -plane they like. We will indicate their respective orientations by the angles φ_a and φ_b , respectively, using $\varphi = 0$ for the positive z -axis. The atoms traverse a field-free region before entering the respective Stern-Gerlach devices, and so there is no change to their spin orientation.⁹⁴ That is, the Heisenberg spin vector operators $\hat{S}_a$ and $\hat{S}_b$ are constant over time.

Suppose now that Alice chooses to set her device to a specific angle φ_a . Due to the spherical symmetry of the state $|\psi_S\rangle_{ab}$, the value of φ_a does not matter, so we can conveniently use $\varphi_a = 0$. What should Alice expect? The two possible outcomes of her experiment are that the atom comes out in the upper or in the lower channel of her device, and symmetry considerations alone are sufficient to establish that the probabilities for these two possible outcomes must be equal. We can, however, also compute them via the Born rule:

$$\Pr(\uparrow, z) = \text{tr}(P_{|\uparrow, z\rangle_a} \rho_a) = \frac{1}{2} \text{tr}(P_{|\uparrow, z\rangle_a}) = \frac{1}{2} = \Pr(\downarrow, z), \quad (3.108)$$

where $\rho_a = \text{tr}_b(P_{|\psi_S\rangle_{ab}}) = \frac{1}{2} \hat{1}_a$ is the reduced state of the atom on Alice's side, which we derived in Eq. (3.107).

Having found her atom in the upper channel, how should Alice update the system's quantum state? The rule of Eq. (3.47) provides a clear answer:

$$|\psi_S\rangle_{ab} \rightarrow |\uparrow, z\rangle_a \otimes |\downarrow, z\rangle_b. \quad (3.109)$$

That is, the theory of QM tells Alice to update the state of the joint system, which includes her atom (on which she has performed a measurement) *and* Bob's, to be in a product state in which both atoms have a fixed and opposite spin orientation. For Bob's atom considered by itself, Alice can consequently update her state ascription to

$$\rho'_b = |\downarrow, z\rangle_b \langle\downarrow, z|, \quad (3.110)$$

where the prime in ρ'_b refers to Alice's state ascription of Bob's atom after her measurement.

Alice can thus use her local knowledge about her outcome to predict a far-away outcome of Bob's experiment with certainty: If he also chooses to measure along the z -direction, he will find the atom in the lower channel. Alice can make probabilistic predictions for Bob's outcome for other orientations as well: if his apparatus is

⁹⁴ For the following considerations, we may assume that the spreading of the atoms' wave packets is negligible while the atoms traverse the field-free region, similar as in our discussion of a homogeneous magnetic field between two Stern-Gerlach devices (Sect. 3.3.3.3). The corresponding Heisenberg equations of motion are then given by Eqs. (3.80) and (3.81) with $\vec{B} = 0$.

rotated away from the positive z -axis by an angle φ in the xz -plane,⁹⁵ then Alice can calculate the probability for Bob to find his atom in the upper channel via

$$\Pr(\uparrow, \varphi) = |\langle \uparrow, \varphi | \downarrow, z \rangle|^2 = \cos^2 \left(\frac{\pi - \varphi}{2} \right) = \sin^2 \left(\frac{\varphi}{2} \right). \quad (3.111)$$

Does this mean that Alice somehow influences or controls Bob's atom? Is there, to use Einstein's oft-quoted expression, "spooky action at a distance"?

The answer is no, and decidedly so. To be fair, the experiment envisaged here does not really pertain to that question anyway: The actual atoms move relatively slowly, and so Alice could signal her prediction to Bob before he chooses a measurement direction, or she could exert some other causal influence on his lab while Bob's atom is under way. But entangled pairs need not be made of slowly moving atoms—in fact, it is technically much easier to create entangled photon pairs, whose individual parts move at the speed of light. With photons, it is possible for Alice and Bob to set their measurement directions while the photons are already in flight, and such that their respective events of setting directions are space-like separated, so that they cannot possibly causally influence one another. Still, once Alice has measured, she can predict with certainty what Bob will find if he chooses a parallel or antiparallel direction, and she can make specific probabilistic predictions for other settings as well. The main point that underlies Einstein et al.'s (1935) claim that quantum mechanics must be incomplete is the (indeed puzzling) fact that Alice is *free to choose* any spin direction on her side and thereby to determine a direction on Bob's side for which she can predict the outcome of Bob's measurement with certainty—if he really chooses to measure the spin along that direction.⁹⁶

What gives Alice this predictive power? What in nature is the basis for her knowledge of what happens in Bob's lab once he measures, which cannot be causally influenced by what happens in hers? The answer given by QM is simple: The basis for her predictions is her previous knowledge of the state of the entangled pair, according to the calculations just shown. What happened in her lab has not influenced Bob's particle, but her knowledge of her local outcome puts her in a position to predict what Bob will find.

Instantaneous updates about distant happenings are a completely normal affair—if you find your lost keys in your deep coat pockets in your office, you can predict instantaneously that they are not at home on your kitchen table. While searching, you may have held that there was a 50% chance of the keys being at home. The "collapse" of that probability value to zero immediately upon your finding the keys is not a causal process, but just your epistemic bookkeeping. The situation is indeed

⁹⁵ The observable measured by Bob's apparatus is then the spin along the direction $\vec{n}_b = \cos \varphi_b \vec{e}_z + \sin \varphi_b \vec{e}_x$, that is, $\vec{n}_b \cdot \hat{S} = \cos \varphi_b \hat{S}_z + \sin \varphi_b \hat{S}_x$.

⁹⁶ It was clear to Einstein et al. that Alice cannot simultaneously predict the outcome of several complementary observables. But the point was that it is completely up to her to choose for which direction she will be able to predict the outcome with certainty.

very similar to the epistemic bookkeeping for the position or momentum of a free particle that we discussed in the context of the spreading of its wave packet (see Eq. (3.60)).⁹⁷

Meanwhile, what should Bob do? Suppose he has set his device to some angle φ_b . If he knows nothing about what happens in Alice's lab, he is in the same position that she was in before she measured—he will predict a 50:50 probability for both possible outcomes (in fact, along any chosen direction, again based on the state's symmetry). But what should Bob do if he knows that Alice has already measured (for example, if his lab is much further away from the source than hers, so that he knows that her particle must have reached her lab before his particle arrives in his lab)? The answer depends in part on what information exactly Bob has available. By the assumptions we are making, he cannot have information about Alice's outcome—he will have to learn about that later, when she tells him. He may or may not have information about Alice's setting, φ_a , depending on whether or not she told him before what she was going to do (and then also lived up to her promise). Suppose he does know φ_a .⁹⁸ Then he knows that Alice must have found either the outcome $|\uparrow, \varphi_a\rangle$ or $|\downarrow, \varphi_a\rangle$. She knows which, but he cannot know that. He knows, however, that both outcomes had equal probability. So he should update from the pure joint state $|\psi_S\rangle_{ab}$ to the density operator for the joint system

$$\rho_{ab} = \frac{1}{2} (|\uparrow, \varphi_a; \downarrow, \varphi_a\rangle_{ab} \langle\uparrow, \varphi_a; \downarrow, \varphi_a| + |\downarrow, \varphi_a; \uparrow, \varphi_a\rangle_{ab} \langle\downarrow, \varphi_a; \uparrow, \varphi_a|). \quad (3.112)$$

For his own particle, this amounts to complete ignorance, which is described by the reduced density operator⁹⁹

$$\rho_b = \text{tr}_a(\rho_{ab}) = \frac{1}{2} (|\downarrow, \varphi_a\rangle_b \langle\downarrow, \varphi_a| + |\uparrow, \varphi_a\rangle_b \langle\uparrow, \varphi_a|) = \frac{1}{2} \hat{1}_b. \quad (3.113)$$

⁹⁷ The difference between the everyday example of the keys and the distributed quantum experiment has to with the principle of complementarity. While we typically imagine that the keys in the above example *are somewhere*, at the very place where you left them (even if you don't remember that place), we should not think that way about the different spin components of one and the same particle. Thinking that both of them have some definite values at the same time, while at the same time accepting Eq. (3.92) as irreducible, would indeed amount to thinking that quantum mechanics must be incomplete.

⁹⁸ Even if Alice told him what she was going to do, he can only check later. His knowledge depends not just on her trustworthiness, but also on the absence of accidents in Alice's lab such as power failures or someone bumping into the apparatus. Bob can have no independent information about Alice's setting or about the success of her experiment before he can exchange information with Alice.

⁹⁹ The reduced density operator for Eq. (3.112) can be obtained in a similar way as in Eq. (3.107) for the singlet state.

Based on his knowledge, Bob will therefore still have to make 50:50 predictions for his own measurement along any direction, including φ_a .

In case Bob does not know about Alice's setting and only knows that she has measured along *some* direction, it seems that he will have even less information regarding the state of the joint system, and the same regarding his own. To reflect that additional ignorance, he could write down the joint, post-measurement state as a probabilistic mixture of all states of the form (3.112):

$$\begin{aligned}\rho_{ab} &= \frac{1}{4\pi} \int_0^{2\pi} d\varphi (|\uparrow, \varphi; \downarrow, \varphi\rangle_{ab} \langle\uparrow, \varphi; \downarrow, \varphi| + |\downarrow, \varphi; \uparrow, \varphi\rangle_{ab} \langle\downarrow, \varphi; \uparrow, \varphi|) \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\varphi (|\uparrow, \varphi; \downarrow, \varphi\rangle_{ab} \langle\uparrow, \varphi; \downarrow, \varphi|),\end{aligned}\tag{3.114}$$

where the last equality holds since $|\downarrow, \varphi\rangle$ corresponds to $|\uparrow, \varphi + \pi\rangle$ and the integration is over a full angular range of 2π . It can be easily checked that the reduced state of Bob's system is, again, the completely mixed state, using Eq. (3.105):

$$\begin{aligned}\rho_b &= \frac{1}{2\pi} \int_0^{2\pi} d\varphi (|\langle\uparrow, z|\uparrow, \varphi\rangle_a|^2 |\downarrow, \varphi\rangle_b \langle\downarrow, \varphi| + |\langle\downarrow, z|\uparrow, \varphi\rangle_a|^2 |\downarrow, \varphi\rangle_b \langle\downarrow, \varphi|) \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\varphi \left(\cos^2 \frac{\varphi}{2} |\downarrow, \varphi\rangle_b \langle\downarrow, \varphi| + \sin^2 \frac{\varphi}{2} |\downarrow, \varphi\rangle_b \langle\downarrow, \varphi| \right) \\ &= \frac{1}{2\pi} \int_0^{2\pi} d\varphi (|\downarrow, \varphi\rangle_b \langle\downarrow, \varphi|) \\ &= \frac{1}{2\pi} \int_0^\pi d\varphi (|\downarrow, \varphi\rangle_b \langle\downarrow, \varphi| + |\uparrow, \varphi\rangle_b \langle\uparrow, \varphi|) \\ &= \frac{1}{2\pi} \int_0^\pi d\varphi \hat{1}_b = \frac{1}{2} \hat{1}_b,\end{aligned}\tag{3.115}$$

where in the second last step we have used Eq. (3.90). The upshot is that unless Bob knows both Alice's choice of setting and her result, his reduced state will be the fully mixed one. Alice's measurement really has no influence on Bob's particle.

It is tempting at this point to compare Alice's and Bob's updates for the description of Bob's system, and declare that Alice, whose update is more specific, has got it right, whereas Bob lacks information. But this would be missing an important point. Both descriptions are perfectly valid; they are really the best that an agent in Alice's or in Bob's position could do. And both provide proper guidance for whatever they choose to do. Bob is still free to set his measurement direction φ_b any way he likes; he knows that in any case, each of the two possible outcomes has a 50:50 chance of occurring. He simply has to update his state ascription

after measurement to be either $|\uparrow, \varphi_b\rangle$ or $|\downarrow, \varphi_b\rangle$. Alice can neither control his selection, nor does she influence his outcome. Even if she can predict Bob's outcome with certainty (which she can for his free choice $\varphi_b = \varphi_a$ or $\varphi_b = \varphi_a + \pi$), she only knows about a perfect correlation one side of which is her own random outcome, which she could not influence at all. No causal influence is exerted, and all interactions and updates are local. When Alice and Bob meet later to compare the statistics of many runs of their experiment, they will be in a position jointly to validate that the source was in fact correctly described by the joint state $|\psi_S\rangle_{ab}$. That description of a (locally produced) particle pair already includes all the correlations they can find.

This account of what is going on in terms of standard QM is straightforward. It presupposes, however, that Alice and Bob both accept indeterminism of two distinct kinds, both agential and natural. Put differently, the account only presupposes that Alice and Bob have retained their everyday assumption that they are free agents facing an open future. The future must be really open with respect to φ_a and φ_b for Alice and Bob to freely select their settings, and the outcome of their respective local experiments must be really open before one of the two possible outcomes is realised. The resulting picture is encouragingly simple: No non-local influences are at play on that account, and everything that happens is based on the local realisation of open possibilities.

Deriving Bell's Theorem As we said, not all interpreters of QM are happy with this account, triggering the search for alternatives. At its simplest,¹⁰⁰ an alternative account would posit a source of particle pairs that issues *different types* of such pairs, discerned by a hidden variable λ , such that for each type, the outcome of spin projection measurements for both particles of the pair is determined in advance. On such an alternative account, the particles are thus given *instruction sets* at the source, as one might say (see Mermin, 1981, 1990). Consider now three directions that Alice and Bob might choose from, corresponding to the angles α , β , and γ . By the properties of $|\psi_S\rangle_{12}$, if Alice and Bob choose the same direction, they must find opposite outcomes. This implies that the instructions must come in eight different types: If Alice's particle carries instructions $\alpha\uparrow\beta\downarrow\gamma\uparrow$, then Bob's particle must carry instructions $\alpha\downarrow\beta\uparrow\gamma\downarrow$, and so on. Bell's surprising result is that this seemingly innocuous assumption about instructions already suffices to derive *empirical predictions that deviate from those of QM*.

The proof goes as follows. Suppose that N particle pairs are created, each of them being of one of the eight types. Denoting the number of pairs in each sub-population by N_i , we have $\sum_{i=1}^8 N_i = N$. Let us fix N_1 to count the pairs with the instruction "all $\uparrow$ for Alice" (and hence, "all $\downarrow$ for Bob"), N_2 for "A: $\uparrow\uparrow\downarrow$ ", N_3 for "A: $\uparrow\downarrow\uparrow$ ", and so on in binary order to N_8 for Alice's " $\downarrow\downarrow\downarrow$ ". Note that each $N_i \geq 0$. This

¹⁰⁰ We follow a standard line of simplifying the original result by Bell (1964); see, e.g., Sakurai and Napolitano (2020, Ch. 3.10). For more general considerations, including a discussion of the more commonly used CHSH (Clauser-Horne-Shimony-Holt) inequality that is based on four rather than three measurement directions, we refer again to Myrvold et al. (2024).

implies, in particular, that

$$N_3 + N_4 \leq (N_2 + N_4) + (N_3 + N_7). \quad (3.116)$$

Now for a particle pair drawn randomly from the population of our N pairs, the probability of Alice finding “ $\uparrow$ ” upon selecting α and Bob finding “ $\uparrow$ ” upon selecting β is

$$\Pr(\uparrow, \alpha; \uparrow, \beta) = \frac{N_3 + N_4}{N}. \quad (3.117)$$

Similarly, we have

$$\Pr(\uparrow, \alpha; \uparrow, \gamma) = \frac{N_2 + N_4}{N}; \quad \Pr(\uparrow, \gamma; \uparrow, \beta) = \frac{N_3 + N_7}{N}. \quad (3.118)$$

Substituting in Eq. (3.116) then yields

$$\Pr(\uparrow, \alpha; \uparrow, \beta) \leq \Pr(\uparrow, \alpha; \uparrow, \gamma) + \Pr(\uparrow, \gamma; \uparrow, \beta). \quad (3.119)$$

This constraint, Bell’s inequality, is not fulfilled by quantum-mechanical expectation values for all directions α , β , and γ . To see this, we can choose $\alpha = 0$, $\beta = \pi/2$, and $\gamma = \pi/4$. Via Eq. (3.111), we can compute (to three significant digits)

$$\begin{aligned} \Pr(\uparrow, \alpha; \uparrow, \beta) &= |\langle \uparrow, \alpha; \uparrow, \beta | \Psi_S \rangle_{12}|^2 \\ &= \frac{1}{2} |\langle \uparrow, \alpha | \uparrow, \alpha \rangle_1|^2 |\langle \uparrow, \beta | \downarrow, \alpha \rangle_2|^2 \\ &= \frac{1}{2} \times 1 \times \sin^2\left(\frac{\pi}{4}\right) = \frac{1}{4} = 0.250; \end{aligned} \quad (3.120)$$

$$\Pr(\uparrow, \alpha; \uparrow, \gamma) = \Pr(\uparrow, \gamma; \uparrow, \beta) = \frac{1}{2} \sin^2\left(\frac{\pi}{8}\right) = 0.073. \quad (3.121)$$

Substituting in Eq. (3.119) then yields an obvious falsehood:

$$0.250 \leq 0.073 + 0.073 = 0.146. \quad (3.122)$$

Starting with our assumption that particle pairs come in different (hidden) types such that they supply instructions for Alice’s and Bob’s outcomes, we have thus arrived at a contradiction with the predictions of QM. And as we already said, these predictions have been vindicated spectacularly successfully by now. This means

that the existence of hidden types of particle pairs that influence outcomes in the way here assumed has been ruled out on empirical grounds.¹⁰¹

As Bell himself emphasised, this does not mean that hidden variables are impossible. He knew Bohm's theory well. But the range of hidden variable theories that can reproduce the predictions of QM is severely constrained by his result. Certainly a successful hidden variable theory cannot live up to the deterministic ideal of a theory that we described in Sect. 3.4.2 above. For example, Bohm's theory fails this ideal because it does not describe a quantum system as a mosaic of independently variable and locally interacting parts.

The Question of “Local Realism” What we have said above, including the derivation of the contradiction in Eq. (3.122), should be uncontroversial. Controversy starts immediately, however, when we move to questions of the *philosophical interpretation* of what has been shown. For example, it is often said that what has been shown is that “quantum mechanics is non-local”. This is at least severely misleading. Bell's result does not concern the theory of QM (which describes correlations between distant events, to be sure, but based on local interactions), but rather, possible or impossible alternatives to that well-established theory. Bohm's theory is one such alternative, and it *is* non-local—but Bohmian mechanics is not QM, even though its predictions for non-relativistic phenomena agree with those of QM. This does establish that many of the phenomena to which QM applies can be described by a non-local theory. But that is hardly surprising—local theories are more tightly constrained than non-local ones, so the latter set is simply larger. The issue about which the controversy revolves appears to be whether *nature is non-local* (in some sense that has to be made sufficiently precise), i.e., whether “local realism” has been shown to be false. As a matter of fact, there is a fairly widely accepted usage of “local realism” in physics, according to which any quantum system that exhibits Bell-type (non-classical) correlations “violates local realism”.¹⁰² In our view, however, it is better to avoid this way of talking. As we hope to have shown, standard QM offers a perfectly local and realistic description of how Alice comes by her non-local update for what Bob will find. And when they come together to establish the correlations giving rise to the failure of Eq. (3.119), that is again a local affair.

By this we do not mean to say that the correlations or other features of entangled quantum systems do not seem puzzling. Entangled quantum systems allow for

¹⁰¹ The simplifications made in our derivation, which are for purposes of exposition only, can be removed without influencing the verdict. In fact, current empirical results such as those mentioned in Footnote 83 mostly concern the CHSH inequality that works with four rather than three directions (two each for Alice and for Bob). It is not necessary to assume that instructions influence results deterministically. The crucial assumptions that go into a derivation of the CHSH inequality only concern the factorisability of probability distributions for outcomes. We again refer to Myrvold et al. (2024) for a detailed account.

¹⁰² An internet search for the expression “violation of local realism” provides a long list of references. See, e.g., Abellán et al. (2018), Gröblacher et al. (2007), Gachechiladze et al. (2016), and Gühne et al. (2005).

unfamiliar ways of transmitting information (see Sect. 3.4.4.1 for the protocol of dense coding), and we have just seen that they can exhibit non-local correlations of a form that is deeply unfamiliar. At least, most of us have not yet familiarised ourselves with this new kind of resource provided by nature. Instantaneous non-local correlations are completely familiar to us—recall from our above example that if you find your keys, you gain instantaneous information about where they are not. But such familiar epistemic updates allow for a simple classical background story of the origin of the correlations, and *that* is what Bell’s theorem has shown to be excluded in the quantum case. Long familiarity with the mentioned images of what a good theory must be, and perhaps other, less theoretical aspects of our experience have surely conditioned us such that we do not expect correlations of that strength. And indeed, one cannot reproduce such quantum-strength correlations by a mechanism that one can build from parts that obey classical laws, such as wires, switches and standard light-bulbs (in fact, that is one way of explaining the content of Bell’s theorem; see Mermin 1981).

But then, one *can* reproduce such correlations by a mechanism that uses, e.g., entangled photon pairs; the required hardware can be ordered online.¹⁰³ We expect that at least some readers will react to this by thinking, “But isn’t this cheating? That is not a mechanism!” Well, yes, not a mechanism of the tinkertoy variety. But rather than looking for such a mechanism that explains how on earth nature could be this way, which has just been shown to be impossible, it seems better to accept that this is how things are on earth, really. Entanglement is a resource that we are learning to control ever better, and in this respect, it is no different from other natural resources that we have learned to harness. Think, for example, about superconductivity, which is crucial for making modern particle accelerators work. “How could a current flow through a conductor without any resistance?” In some relevant sense, there is no answer to this question. We have a good model, yes. But not “a mechanism”. Rather, we have come to accept that this is an effect that occurs in nature—and we can meanwhile build cool stuff based on that insight. With respect to quantum correlations, we may be approaching a similar point of technological development at which bewilderment slowly gives way to familiarity due to real-world applications.

3.4.4 *Quantum Information Theory: Bell-type States as a Resource*

The quantum correlations exhibited by the Bell singlet state and the philosophical interpretation of these correlations (including popular terms like spooky action at a distance, etc.) have inspired the search for novel applications, for example, regarding

¹⁰³ As of early 2024, one can buy turn-key, high-brightness correlated photon pair sources for around the price of a good motorcycle.

the possibility of communicating with the help of elementary quantum systems. The entanglement of the Bell singlet state has thereby acquired a paradigmatic role for the entire field of quantum information. In Sect. 3.4.4.1 we describe the general background and two popular protocols, dense coding and quantum teleportation. Sections 3.4.4.2 and 3.4.4.3 together with Sect. 3.5.4 then provide more technical background that prepares the discussion of advanced quantum applications in Sect. 5.4. These sections could be skipped on a first reading.

3.4.4.1 Multiple Qubits

Based on our description of the Bell singlet state (Eq. (3.102)) and some of its properties, we now provide some more background on quantum information theory, expanding our description of single qubits from Sect. 3.3.3.2 to entangled multipartite systems such as the Bell state and beyond.¹⁰⁴ This will set the scene for discussions of several protocols known in quantum information. These protocols are often cast in the framework of parties, or agents, who perform a series of actions on quantum systems that they control in their laboratories, including unitary operations (e.g., through the application of electromagnetic fields) and measurements. Building on this background, in Chap. 5 we will go one step further by including models of agents into the description.

The Bell singlet state $|\psi_S\rangle_{12} \in \mathcal{H}_1 \otimes \mathcal{H}_2$ from Eq. (3.102) in our qubit notation of Sect. 3.3.3.2 takes the form

$$|\Psi_S\rangle_{12} = \frac{1}{\sqrt{2}} (|0\rangle_1 \otimes |1\rangle_2 - |1\rangle_1 \otimes |0\rangle_2) . \quad (3.123)$$

It is a simultaneous eigenstate of two observables and satisfies the eigenvalue equations

$$\begin{aligned} (\sigma_z^{(1)} \otimes \sigma_z^{(2)})|\Psi_S\rangle_{12} &= -|\Psi_S\rangle_{12} , \\ (\sigma_x^{(1)} \otimes \sigma_x^{(2)})|\Psi_S\rangle_{12} &= -|\Psi_S\rangle_{12} , \end{aligned} \quad (3.124)$$

which can easily be checked. The two observables $\sigma_z^{(1)} \otimes \sigma_z^{(2)}$ and $\sigma_x^{(1)} \otimes \sigma_x^{(2)}$ have in fact four different eigenstates that are mutually orthogonal and span the entire Hilbert space $\mathcal{H}_{12} = \mathcal{H}_1 \otimes \mathcal{H}_2$. These eigenstates form a basis of maximally entangled states, which is called the *Bell basis* of $\mathcal{H}_{12}$. Here and in some of the following examples we use a simplified notation, dropping both the tensor product symbol and the particle indices, and we adapt the notation $|\Psi^-\rangle = |\Psi_S\rangle$ for the singlet state:

¹⁰⁴ A large body of research in quantum information theory deals with the mathematical aspects of entanglement, including, e.g., its classification and quantification. For excellent reviews of this research see Horodecki et al. (2009) and Bengtsson and Życzkowski (2017).

$$\begin{aligned}
|\Phi^+\rangle &= \frac{1}{\sqrt{2}} (|00\rangle + |11\rangle) , \\
|\Phi^-\rangle &= \frac{1}{\sqrt{2}} (|00\rangle - |11\rangle) , \\
|\Psi^+\rangle &= \frac{1}{\sqrt{2}} (|01\rangle + |10\rangle) , \\
|\Psi^-\rangle &= \frac{1}{\sqrt{2}} (|01\rangle - |10\rangle) .
\end{aligned} \tag{3.125}$$

The Bell states play a central role in many basic applications in quantum information, among which are device-independent randomness (see Sect. 3.5.3), quantum teleportation (Bennett et al., 1993), dense coding (Bennett & Wiesner, 1992), and the quantum repeater (Briegel et al., 1998). In these applications an agent—which can be a human agent, but need not be, as these basic applications can be fully automated—interacts with the system by a sequence of interventions that correspond, e.g., to unitary processes or measurements.

Remarkably, Bell states can be transformed into each other by applying a *local* unitary operation on one qubit, equivalent to a change of basis of the local Hilbert space. By operating locally on a single qubit, an agent can thus rotate the state of the joint two-qubit system into four different orthogonal states. For example, consider the initial situation where two agents, Alice and Bob, located in two spatially separated laboratories, share one of the entangled states of Eq. (3.125), say, $|\Phi^+\rangle_{ab}$. If Bob applies the unitary operation $U_2 = -i\sigma_y$ to his qubit, he will transform the shared state into $|\Psi^-\rangle_{ab}$.¹⁰⁵ More generally, if Bob applies one of the four unitary operations (rotations) $U_0 = \sigma_0 \equiv \hat{1}$, $U_1 = \sigma_1 \equiv \sigma_x$, $U_2 = -i\sigma_2 \equiv -i\sigma_y = \sigma_x\sigma_z$, $U_3 = \sigma_3 \equiv \sigma_z$, he can transform the shared state $|\Phi^+\rangle_{ab}$ into any one of the four Bell states, viz.,¹⁰⁶

$$\begin{aligned}
|\Phi^+\rangle_{ab} &= \hat{1}^{(b)} |\Phi^+\rangle_{ab} , \\
|\Phi^-\rangle_{ab} &= \sigma_z^{(b)} |\Phi^+\rangle_{ab} , \\
|\Psi^+\rangle_{ab} &= \sigma_x^{(b)} |\Phi^+\rangle_{ab} , \\
|\Psi^-\rangle_{ab} &= -i\sigma_y^{(b)} |\Phi^+\rangle_{ab} .
\end{aligned} \tag{3.126}$$

¹⁰⁵ When the qubit corresponds to an atomic spin, such a unitary rotation can be generated by applying a magnetic field in the y -direction; see Eq. (3.82) and the analysis surrounding it. To realise the spin rotation $U_2 = -i\sigma_y$ the B -field has to be applied for the time $t = \pi/\omega$, which rotates $\hat{S}_z(0)$ into $\hat{S}_z(t) = -\hat{S}_z(0)$, as can be seen via Eq. (3.85). For other physical realisations, unitary rotations may also be generated in other ways, for example, by laser pulses when qubits are encoded in the internal states of trapped ions.

¹⁰⁶ Here we use a superscript “(b)” on the Pauli matrices to indicate that these unitary operations are applied to the second qubit, which is controlled by Bob.

The fact that Bell states can be transformed into each other by simple local rotations also means that their entanglement properties are the same (Linden & Popescu, 1998; Vidal, 2000). Since the Bell states are identical up to local unitary rotations, one often picks just one representative of these states, e.g., $|\Phi^+\rangle$.

The relations of Eq.(3.126) are exploited in various quantum information applications, in particular dense coding and teleportation.

Dense Coding For dense coding (Bennett & Wiesner, 1992), Alice and Bob use a shared entangled Bell state as a resource, together with Eq. (3.126), to communicate two bits of classical information by exchanging a single qubit. After applying one of the rotations U_μ to his qubit b , Bob sends this qubit to Alice, who can then measure the state of the joint system of qubits a and b in her laboratory. The orthogonality of the Bell states will allow her to find out, with certainty, which of the four operations Bob applied to his qubit.¹⁰⁷ While the direct, practical advantage of this protocol for the transmission of classical information is limited, its perhaps most remarkable aspect is foundational and connects to our discussion of “local realism” at the end of Sect. 3.4.3. For, as the protocol demonstrates, an agent (Bob) who employs only local physical interactions with the qubit held in his laboratory, nevertheless can have access to a four-dimensional Hilbert space that is constituted by both particles, if they are Bell-type entangled, even if the particles are located far away from each other. This access is operationally verifiable. It represents a notion of non-locality that has no ontological commitment to “local realism” but still a clear operational interpretation in quantum communication applications.

Quantum Teleportation For quantum teleportation (Bennett et al., 1993), Alice and Bob use a shared entangled Bell pair, say, the Bell state $|\Phi^+\rangle_{ab}$ as a resource, as in dense coding, but no qubit is exchanged. That is, qubits a and b are always kept by Alice and Bob, respectively. In addition to qubit a , Alice has another qubit, labelled a' , which is in a state $|\psi\rangle_{a'} = \alpha|0\rangle_{a'} + \beta|1\rangle_{a'}$ that Alice may not know.¹⁰⁸ The protocol of quantum teleportation allows Alice and Bob to transfer the quantum state that is carried by Alice’s qubit a' to Bob’s qubit b , without the need of sending any quantum particle that carries the quantum information $|\psi\rangle$ through the physical channel that connects Alice’s and Bob’s laboratory, just by exploiting the quantum correlations of their shared Bell pair.

The protocol can be divided into three steps (i)–(iii). At the beginning of the protocol, the state of the three particles a' , a , and b has the form $|\psi\rangle_{a'} \otimes |\Phi^+\rangle_{ab}$.

¹⁰⁷ Technically, she may apply a Bell measurement of the state of qubits a and b , corresponding to the measurement of a joint observable $O_{\text{Bell}} = \mu_0|\Phi^+\rangle_{ab}\langle\Phi^+| + \mu_1|\Phi^-\rangle_{ab}\langle\Phi^-| + \mu_2|\Psi^+\rangle_{ab}\langle\Psi^+| + \mu_3|\Psi^-\rangle_{ab}\langle\Psi^-|$, with non-degenerate eigenvalues μ_j , $j = 0, \dots, 3$. Equivalently, she can measure the two observables $\sigma_z^{(a)} \otimes \sigma_z^{(b)}$ and $\sigma_x^{(a)} \otimes \sigma_x^{(b)}$, whose pairwise combinations of measured eigenvalues ($++$, $+-$, $-+$, or $--$) uniquely identify the different Bell states as corresponding eigenstates.

¹⁰⁸ For example, the qubit a' might have been prepared by some third party and given to Alice, without telling her in which state it was prepared. This is why one sometimes says that Alice can teleport an “unknown quantum state” to Bob.

For the description of the protocol, it is useful to rewrite that state in the following way:

$$\begin{aligned}
 (\alpha|0\rangle_{a'} + \beta|1\rangle_{a'}) |\Phi^+\rangle_{ab} = & \frac{1}{2} |\Phi^+\rangle_{a'a} (\alpha|0\rangle_b + \beta|1\rangle_b) \\
 & + \frac{1}{2} |\Phi^-\rangle_{a'a} (\alpha|0\rangle_b - \beta|1\rangle_b) \\
 & + \frac{1}{2} |\Psi^+\rangle_{a'a} (\alpha|1\rangle_b + \beta|0\rangle_b) \\
 & + \frac{1}{2} |\Psi^-\rangle_{a'a} (\alpha|1\rangle_b - \beta|0\rangle_b) .
 \end{aligned} \tag{3.127}$$

This can be easily verified by inserting the explicit expressions (Eq. (3.125)) for the Bell states into the corresponding terms appearing in Eq. (3.127). In the first step, (i) Alice applies a Bell measurement (see Footnote 107) on the local pair of qubits a' and a in her lab. This measurement will project the qubits' joint quantum state into one of the four Bell states of Eq. (3.125) with equal probability of $1/4$. In the next step, (ii) Alice communicates the result of her measurement to Bob, which requires two bits of classical information. In the last step, (iii) Bob will then apply one of the four unitaries $U_\mu^{(b)}$ to his qubit, according to Eq. (3.127) and the explicit expressions for $U_\mu^{(b)}$ specified before Eq. (3.126). The final state of this protocol is then a tensor product of a Bell pair held locally by Alice and the state $|\psi\rangle_b = \alpha|0\rangle_b + \beta|1\rangle_b$, initially carried by Alice's particle a' , now on Bob's particle b . For example, consider the case where Alice's Bell measurement projects her particles into the state $|\Psi^+\rangle_{a'a}$, which corresponds to the third line of Eq. (3.127). At this point, the properly renormalised state of the total system is given by $|\Psi^+\rangle_{a'a} \otimes (\alpha|1\rangle_b + \beta|0\rangle_b)$. As one can read off from this expression, the state of Bob's particle has the form $(\alpha|1\rangle_b + \beta|0\rangle_b) = \sigma_x^{(b)} (\alpha|0\rangle_b + \beta|1\rangle_b)$. If Alice tells Bob which result she obtained, Bob can then apply the correction operator $\sigma_x^{(b)}$ to his qubit. As a result, the same logical information that was initially carried by Alice's particle a' is now carried by Bob's particle b .

The entire protocol can be conveniently summarised by the following formula:

$$\begin{aligned}
 |\psi\rangle_{a'} |\Phi^+\rangle_{ab} = & \frac{1}{2} \left(|\Phi^+\rangle_{a'a} U_0^{(b)} + |\Phi^-\rangle_{a'a} U_3^{(b)} \right. \\
 & \left. + |\Psi^+\rangle_{a'a} U_1^{(b)} + |\Psi^-\rangle_{a'a} U_2^{(b)} \right) |\psi\rangle_b ,
 \end{aligned} \tag{3.128}$$

from which the steps (i)–(iii) can be read off. The $|\psi\rangle$ that appears on both sides of the equation describes the same logical qubit, but that state is ascribed to different physical particles that carry this information.

It should be stressed that teleportation does not *copy* quantum information in any sense. For, after the Bell measurement, neither particle a' , nor particle a , nor both of them together contain any information about the logical qubit $|\psi\rangle$. Furthermore,

there is no instantaneous (or “superluminal”) transfer of quantum information. As long as Bob does not know the classical result of Alice’s measurement, he has no reason to assign any state to his particle b other than the completely mixed state $\rho_b = \frac{1}{4} \sum_{\mu} U_{\mu}^{(b)} |\psi\rangle_b \langle\psi| U_{\mu}^{\dagger(b)} = \frac{1}{2} \hat{1}_b$.

3.4.4.2 Stabilizer States

The state $|\Phi^+\rangle$ is an example of a more general class of states called *stabilizer states*. A stabilizer state of N qubits is defined as the common eigenstate with eigenvalue $+1$ of a set of N independent commuting Pauli observables $M_k \in \mathcal{P}^{\otimes N}$, $k = 1, \dots, N$ (Nielsen & Chuang, 2010, Ch. 10.5). Here $\mathcal{P}$ denotes the Pauli group of a single qubit, which is generated by the Pauli matrices $\sigma_x, \sigma_y, \sigma_z$ introduced in Eq. (3.72). It contains these matrices and all of their products, which includes the identity matrix $\sigma_0 \equiv \hat{1}$, together with additional phase factors ± 1 and $\pm i$. Formally, the Pauli group can be written as the span of the three Pauli matrices $\sigma_x, \sigma_y, \sigma_z$ as generators:

$$\mathcal{P} = \langle \sigma_x, \sigma_y, \sigma_z \rangle = \{\pm 1, \pm i\} \times \{\hat{1}, \sigma_x, \sigma_y, \sigma_z\}. \quad (3.129)$$

A stabilizer state $|\psi\rangle$ of N qubits thus satisfies a set of eigenvalue equations

$$M_k |\psi\rangle = |\psi\rangle \quad (3.130)$$

for $k = 1, \dots, N$. The stabilizer operators M_k commute and have the explicit form

$$M_k = \prod_{n=1}^N \tilde{\sigma}_{\mu_n(k)}^{(n)}, \quad (3.131)$$

where each of the factors $\tilde{\sigma}_{\mu_n(k)}^{(n)}$, $\mu_n(k) \in \{0, 1, 2, 3\}$, denotes one of the four Pauli matrices (modulo possible phase factors $\{\pm 1, \pm i\}$, that is, $\tilde{\sigma}_{\mu_n(k)}^{(n)} \in \mathcal{P}$), the superscript n identifies the qubit on which they act, and one requires that the group generated by the M_k not include $-\hat{1}$. Note that in the multi-partite Hilbert space of an N -qubit system one often identifies the operator $\sigma_{\mu_n}^{(n)}$ that acts on qubit n with its trivial extension to the operator $\sigma_{\mu_n}^{(n)} \otimes_{n' \neq n} \hat{1}^{(n')}$ that acts as the identity on the other qubits $n' \neq n$. As the stabilizer operators commute, $|\psi\rangle$ is also an eigenstate of any product of different operators M_k , i.e., of any operator from the (abelian) stabilizer group, $S_N = \langle M_1, M_2, \dots, M_N \rangle = \{M_1^{s_1} M_2^{s_2} \dots M_N^{s_N} | s_1, s_2, \dots, s_N \in \{0, 1\}\}$. The stabilizer group is also referred to as *the stabilizer*, and the operators $M_1, M_2, \dots, M_N$ as *generators* of S_N . For notational simplicity, one often uses the shorthand notation

$$I_n = \sigma_0^{(n)}; \quad X_n = \sigma_x^{(n)}; \quad Y_n = \sigma_y^{(n)}; \quad Z_n = \sigma_z^{(n)}. \quad (3.132)$$

With this notation, the Bell state $|\Phi^+\rangle \equiv |\Phi^+\rangle_{12}$, for example, is given as the unique solution of the two eigenvalue equations

$$\begin{aligned} X_1 X_2 |\Phi^+\rangle &= |\Phi^+\rangle, \\ Z_1 Z_2 |\Phi^+\rangle &= |\Phi^+\rangle. \end{aligned} \quad (3.133)$$

Furthermore, one often drops the index n or the identity operator I_n when there is no danger of confusion. The stabilizer of the Bell state $|\Phi^+\rangle$ can then be written as $S_{|\Phi^+\rangle} = \langle X_1 X_2, Z_1 Z_2 \rangle = \langle XX, ZZ \rangle$ while for the product state $|00\rangle$ the stabilizer can be written as $S_{|00\rangle} = \langle Z_1 I_2, I_1 Z_2 \rangle = \langle ZI, IZ \rangle = \langle Z_1, Z_2 \rangle$.

The strength of the stabilizer formalism is not yet apparent for simple quantum states such as the Bell states. For systems with a larger number N of qubits, however, it is often much simpler to specify the N -qubit quantum state by its stabilizer operators $M_1, \dots, M_N$ than by its explicit expansion in a basis of single-qubit states; we will shortly see examples of this. Furthermore, many applications of quantum information processing involve measurements of individual qubits in an entangled state, as for example in quantum error correction or in measurement-based quantum computation. The inherent randomness of the individual measurement outcomes, an example of which we have already seen in the discussion of Bell's theorem in Sect. 3.4.3, then leads to a branching of the state ascribed to the other, yet unmeasured qubits. To efficiently keep track of these random events and the corresponding updates of the system's quantum state, the stabilizer formalism is very helpful (Gottesman, 1997). It tells an agent how to actively compensate for random measurement outcomes in order to realise an effectively deterministic, unitary evolution, if that is desired.¹⁰⁹ From a broader perspective, the efficient mathematical description of the quantum state of a system and of how it needs to be updated due to uncontrolled interactions with the environment or controlled actions by an agent, is also relevant for the agent's capacity of modelling and simulating the system it is dealing with. Before we give concrete examples, we discuss an important graphical representation of stabilizer states as *graph states* (Hein et al., 2004, 2006).

3.4.4.3 Graph States

Consider a mathematical graph $G = (V, E)$, where V denotes the set of vertices and $E \subseteq V^2$ the set of (undirected) edges.¹¹⁰ A graph state $|G\rangle \in \mathcal{H}^{\otimes N}$ of a system

¹⁰⁹ As we already mentioned, entangled quantum states are nowadays understood as key resources for applications in quantum information and technology. These can either be applications that make direct use of the inherent randomness of local measurement outcomes, as in the case of device-independent randomness generation (see Sect. 3.5.3), or they can be applications where a controller—or, more generally, an agent—interacts with the quantum system in order to stabilize it in a certain state (quantum memory) or to transform it to a desired target state (general quantum computer) in the presence of random events that are the result of measurements or environmental noise.

of $N = |V|$ qubits, each of which is associated with a vertex of the graph G , is defined as the common eigenstate with eigenvalues $+1$ of the set of N correlation observables $K_a \equiv K_a(G)$, that is:

$$K_a|G\rangle = |G\rangle, \quad (3.134)$$

with

$$K_a = X_a \prod_{b \in N_a} Z_b \quad (3.135)$$

for all $a \in V = \{1, 2, \dots, N\}$, where $N_a \equiv \{b \in V | \{a, b\} \in E\}$ denotes the set of neighbours of a , i.e., the set of vertices connected by an edge to vertex a (Raussendorf et al., 2003; Hein et al., 2004). A graph state $|G\rangle$ is thus a stabilizer state for which the stabilizer has a simple graphical description in terms of G . Graph states are a sub-class of stabilizer states. Interestingly, one can show that every stabilizer state is local-unitarily equivalent to some graph state, which means that graph states provide a *normal form* for all stabilizer states (Schlingemann, 2002; Grassl et al., 2002). The formalism of graph states is useful to model situations where an agent applies a series of specific operations (actions) on the system, which consist of measurements of the three Pauli observables and certain unitary operations that transform stabilizer states into other stabilizer states. The effect of these actions on the quantum state can be efficiently described by graph-theoretical transformations on the graph representing the quantum state, together with local unitary so-called by-product operations (or, equivalently, local basis changes). For example, after the measurement of the Pauli observable Z_a of qubit $a \in V$, which has the random outcome $(-1)^s = \pm 1$ for $s \in \{0, 1\}$ (as we will prove in the next paragraph), the resulting state of the system is of the form

$$|s\rangle_a \otimes U_{z,a}^s |G - a\rangle, \quad (3.136)$$

where $|G - a\rangle$ is again a graph state associated with a modified graph in which the vertex a has been removed from the vertex set V , together with all edges from set E which are incident on a . The unitary by-product operator $U_{z,a}^s$ is given by

$$U_{z,a}^s = Z_{N_a}^s \equiv \left(\prod_{b \in N_a} Z_b \right)^s. \quad (3.137)$$

This means, if the result of the z -measurement on qubit a is $+1$ ($s = 0$), then the state of the remaining qubits is the graph state $|G - a\rangle$; while if the result is

¹¹⁰ We consider simple graphs where there are no multiple edges between the same pair of vertices and no edge connects a vertex with itself. See, e.g., Diestel (2017) for an introduction to graph theory.

-1 ($s = 1$), then it is locally unitarily equivalent to $|G - a\rangle$ and the agent needs to apply a z -rotation to every qubit b in the graph-theoretical neighbourhood of a . Transformation rules similar to Eq. (3.137), together with other properties of graph states, are given by Hein et al. (2004, 2006).

Graph states have many interesting properties, one of which we highlight here.¹¹¹ It is straightforward to show that the measurement of any qubit $a \in V$, along any direction $\vec{n}$, has a random outcome. This is equivalent to showing that the expectation value of the spin in any direction vanishes, i.e., $\langle \sigma_{\vec{n}}^{(a)} \rangle = \langle G | \vec{n} \cdot \vec{\sigma}^{(a)} | G \rangle = 0$. Consider first the special case where $\vec{n} = \vec{e}_z$, i.e., $\vec{n} \cdot \vec{\sigma}^{(a)} = \sigma_z^{(a)} = Z_a$. We have

$$\langle G | Z_a | G \rangle = \langle G | Z_a K_a | G \rangle = -\langle G | K_a Z_a | G \rangle = -\langle G | Z_a | G \rangle, \quad (3.138)$$

from which one obtains $\langle G | Z_a | G \rangle = 0$. The first and the third equalities hold since $|G\rangle$ and $\langle G|$ are eigenkets and eigenbras (right and left eigenstates) of K_a , respectively, with eigenvalue $+1$. The second equality holds due to the explicit form of K_a given in Eq. (3.135) and the fact that the different Pauli operators (of the same qubit) anticommute, in particular, $Z_a X_a = -X_a Z_a$. By similar arguments one finds that for $b \in N_a$,

$$\langle G | X_a | G \rangle = \langle G | X_a K_b | G \rangle = -\langle G | K_b X_a | G \rangle = -\langle G | X_a | G \rangle, \quad (3.139)$$

and the same holds if we replace X_a with Y_a in the derivation. Since the spin observable $\sigma_{\vec{n}}^{(a)}$ along an arbitrary direction $\vec{n}$ can be written as a linear combination of X_a , Y_a and Z_a , we also have

$$\langle G | \sigma_{\vec{n}}^{(a)} | G \rangle = 0. \quad (3.140)$$

Correspondingly, the probabilities of the outcomes ± 1 when measuring $\sigma_{\vec{n}}^{(a)}$ are

$$\text{Pr}(\pm, \vec{n}) = \frac{1 \pm \langle \sigma_{\vec{n}}^{(a)} \rangle}{2} = \frac{1}{2}, \quad (3.141)$$

as claimed.¹¹²

¹¹¹ In the following, we consider graph states $|G\rangle$ for which the underlying graph G is *connected*. A graph $G = (V, E)$ is connected if for every two vertices $a, b \in V$ there is a set of vertices $a = a_1, a_2, \dots, a_L = b$ that are connected by edges, i.e., $\{a_l, a_{l+1}\} \in E$ for $l = 1, 2, \dots, L - 1$.

¹¹² A different way to arrive at this result is to show that the reduced quantum state of any qubit a , when the total system is in a (connected) graph state $|G\rangle$, is the completely mixed state. That is, $\rho_a = \text{tr}_{V-a}(|G\rangle\langle G|) = \frac{1}{2}\hat{1}_a$, where $V - a$ denotes the set of all vertices of G with vertex a removed, and tr_{V-a} is the partial trace over all qubits except the qubit at vertex a . The reduced state ρ_a of the total state ρ_V of a set of $|V| = N$ qubits can be obtained from ρ_V by successively tracing over all qubits $b \in V - a$. This can be done by successively applying the formula in Eq. (3.105)

Remarkably, while the outcome of a measurement on any individual qubit alone is completely random, there are strict correlations between the outcomes of certain measurements on a group of different qubits of a graph state. From Eqs. (3.134) and (3.135) one readily obtains $\langle G|K_a|G\rangle = 1$ for the expectation value of all stabilizer observables K_a , which means that

$$\langle K_a \rangle = \langle X_a \prod_{b \in N_a} Z_b \rangle = 1 \quad (3.142)$$

for any $a \in V$. This relation implies strict correlations between the measurement outcomes of specific X and Z observables of different individual qubits.¹¹³ These strict multi-qubit correlations correspond to and generalise the strict correlations that we discussed in the context of the Bell singlet state of Eq. (3.102). Furthermore, similar as for the singlet state one can derive a Bell-type inequality for every (connected) graph state of $N \geq 3$ qubits (Gühne et al., 2005).

The intriguing interplay between indeterminism and entanglement that we noted for the quantum correlations in the Bell states of two qubits thus carries over to the much larger set of graph states of N qubits.

In Fig. 3.6 we show some examples of graph states that are resources for different quantum applications. In Fig. 3.6a we see a generic example of the graph representing the stabilizer operators K_a as defined in Eq. (3.135); the neighbours of vertex a are b , b' , and b'' . In Fig. 3.6b we see the graph-state representation of the Bell state, which corresponds to a graph with two vertices that are connected by an edge (indicated by a line). Such drawings are often used informally to represent two qubits that are entangled, but here the graph has a concise mathematical meaning and represents two eigenvalue equations whose solution describes the quantum state, viz., $Z_1 X_2 |G\rangle = |G\rangle$ and $X_1 Z_2 |G\rangle = |G\rangle$.¹¹⁴ This double role of the graphical formalism is similar to the double role of Feynman diagrams in high-energy physics, which are used both as pictorial diagrams describing the propagation and interaction of elementary particles in space-time, and as formal diagrams representing mathematical integrals that appear in the perturbation expansion of the scattering matrix. In Fig. 3.6c we see the star-shaped graph corresponding to a Greenberger-Horne-Zeilinger (or GHZ) state (Greenberger et al., 1990) of several

and considering the system as a two-partite system, with one part being qubit b and the other part the remaining qubits.

¹¹³ Similar relations can be obtained from any observable which belongs to the stabilizer group associated with $|G\rangle$.

¹¹⁴ The explicit form of the two-qubit graph state is given by $|G\rangle = 1/\sqrt{2}\{|0\rangle_z|0\rangle_x + |1\rangle_z|1\rangle_x\}$, which is the Bell state $|\Phi^+\rangle$, up to a local unitary in the Hilbert space of the second qubit that rotates the $|\cdot\rangle_z$ basis into the $|\cdot\rangle_x$ basis, and *vice versa*. Since all Bell states are equivalent under local unitary rotations, the two-qubit graph state $|G\rangle$ can be considered as a representative of *all* Bell states. Connecting to the discussion around Eq. (3.126), one can write $|\Phi^+\rangle_{12} = U_H^{(2)}|G\rangle$, with the Hadamard unitary rotation $U_H^{(2)} = (\sigma_z^{(2)} + \sigma_x^{(2)})/\sqrt{2}$, and similarly for the other Bell states in Eq. (3.126).

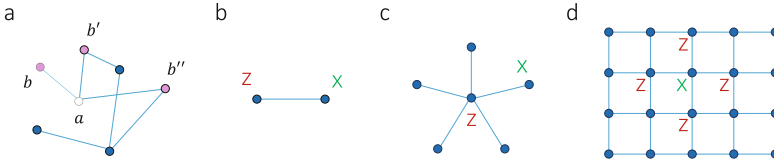


Fig. 3.6 Examples of graph states. (a) Generic case; (b) Bell state; (c) GHZ state; (d) Cluster state

qubits, which also plays an important role in foundational discussions and which will reoccur in the context of quantum error correction in Sect. 3.5.4.3. The graph describing a lattice in Fig. 3.6d represents a 2D cluster state (Briegel & Raussendorf, 2001), which is a universal resource for measurement-based quantum computation (Raussendorf & Briegel, 2001).

The properties of graph states highlight the intriguing interplay of indeterminism and entanglement that is fundamental to quantum mechanics and many of its modern applications. Here we mention three such applications, two of which will be discussed in more detail in Sects. 3.5.3 and 3.5.4 below.

- (1) Device independent randomness generation exploits Bell's theorem for entangled, two-qubit (and larger) graph states to guarantee the irreducible randomness of the bits produced by the measurements. We will discuss this protocol in Sect. 3.5.3.
- (2) Quantum error correction (QEC) deals with the inevitable occurrence of noise in a quantum memory. Noise is always present due to the uncontrolled residual interactions of the environment with the physical qubits that constitute the memory. To protect the memory against noise, QEC encodes the quantum information into entangled states (similar to graph states) of a larger number of physical qubits in such a way that all information resides in the joint entangled state of the quantum memory, while the individual parts (i.e., the reduced local states of the individual qubits) carry no information. If the encoding is well-chosen, the information that is stored in the memory as a whole cannot be destroyed when the environment interacts, locally, with a single qubit. In a sense, quantum error correction exploits indeterminism (which comes with the controlled entanglement of the physical qubits that constitute the quantum memory) to fight indeterminism (which comes with the uncontrolled entanglement of the physical qubits of that quantum memory with the environment). QEC will be further discussed in Sect. 3.5.4.
- (3) Measurement-based (or one-way) quantum computation (MBQC) employs both entanglement and indeterminism for quantum computation. In the standard circuit model of a quantum computer (Nielsen & Chuang, 2010), a quantum computation corresponds to a controlled unitary evolution of a quantum register (i.e., of a set of qubits). This dynamics is realised as a sequence of elementary quantum logic gates, which are generated by one- and two-qubit interactions, similar as in the Boolean circuit model of classical computation. MBQC, in

contrast to that circuit model, works without interactions between the qubits. Instead, any quantum computation corresponds to a sequence of measurements on a specific entangled state that is given as a resource. The one-way quantum computer as proposed by Raussendorf and Briegel (2001) uses the *cluster state*, which is the paradigmatic example of a graph state, as its universal resource.¹¹⁵ The quantum computation is then realised by measuring every qubit of the cluster state in a certain direction, where the choice of that direction may depend on the (random) outcomes of previous measurements. This mode of operation then allows one to simulate any quantum logic circuit. It exploits the quantum correlations, or entanglement, of the cluster state in a similar sense as the entanglement of the Bell state is exploited for quantum teleportation in Sect. 3.4.4.1.

Operating in that way, the one-way quantum computer could be classified as an *agent* in the sense used in computer science (Russell & Norvig, 2022): it receives *percepts* from its environment and influences the environment via *actions*. On this view, the (indeterministic) environment of the agent would be a physical system of $N = |V|$ qubits that are in a highly entangled quantum state. The actions of the agent are measurements of different observables, corresponding to the different directions of the spin of a particle, which can be applied to any of the N qubits of the entangled state. The percepts are the measured eigenvalues ± 1 of the chosen observables that the system returns as the outcomes of the measurements. In a one-way quantum computation protocol, it is scripted (fully specified) which measurement the agent needs to perform on every qubit, given the record of the previous measurement outcomes and the algorithm the agent is supposed to be executing.¹¹⁶ At the same time, the outcomes of the measurements (except for the final readout) are typically fully indeterministic. This allows for another mode of operation of MBQC: With only partial adaptation of the measurement bases, one can exploit the inherent randomness of the protocol for generative modelling and other quantum applications (Miller et al., 2017; Majumder et al., 2024).

In Chap. 5 we will revisit some of the mentioned quantum information applications, including MBQC, in the context of Projective Simulation (PS). There we will describe more advanced scenarios where a learning agent, our PS agent, interacts with a quantum system. The agent then not only executes a given protocol (e.g., a fixed quantum computation or error correction protocol), but also learns to adapt its interactions with the quantum system under some broader objective. This can include changing its measurement strategy, optimising error-correcting codes, or designing new quantum experiments. In Sect. 5.6, we will

¹¹⁵ A brief introduction to cluster states and some of their properties can be found in Briegel (2009).

¹¹⁶ Operating in this way, the one-way quantum computer would be classified as a reflex agent rather than a learning agent. However, one can also think of more general scenarios, where a less narrowly defined task is specified and learning plays a role. An example of such a more general scenario is reported by Smith et al. (2023).

even integrate concepts of quantum physics and quantum information into the design of the PS agent itself.

3.5 Quantum Randomness as a Resource for Sophisticated Agents

In what follows, we will describe some of the current developments of quantum technology that witness the mentioned move from a bewildered “how could nature *possibly be* that way” to “wow, nature *is* that way, let’s use this to make cool stuff”. Similar conceptual shake-ups have happened with the advent of other technologies—in addition to superconductivity mentioned above, think for example about semiconductors, radio waves, or the refrigerator. The turn to broadly adopted technological applications signals a conceptual shift in which a previous source of wonderment is turned into an accepted truth on which agents can base their (business) plans. The move to building quantum technology thus means that those parts of the theory that emerging technology relies on (which include fundamental randomness, superposition, and entanglement) is becoming firmly accepted—in the same way in which the general theory of relativity has become firmly accepted due to its decisive role for satellite technology including our ubiquitous GPS.

In the history of science, we can find several episodes that show how the acceptance of such new frameworks unfolds as a social process. We venture to say it will be similar in quantum theory. Consider, for example, Newton’s theory of gravitation, which posited forces acting at a distance. The notion of a force was taken from everyday experience—lifting and pushing things or pulling on strings, for example. The action of such forces is always a strictly local matter, but gravitation could not be understood like that. Newton himself was greatly puzzled by action at a distance.¹¹⁷ But over time, action at a distance was accepted as something that nature exemplified, and the question was put to rest.¹¹⁸ To mention another episode, consider the principle of the identity of indiscernibles, which states that no two distinct things can be qualitatively exactly alike. Newton’s contemporary (and calculus adversary) Leibniz sought to establish that principle as an *a priori* metaphysical truth. (Roughly, Leibniz’s argument was that there could be no sufficient reason for having *two* such things around, with the plenitude of the world already sufficiently well served by one.) Leibniz had good empirical support for the principle: studying the form of leaves and other biological entities, there was always

¹¹⁷ See Ducheyne (2014) for a recent account that also provides an overview of the discussion in Newton scholarship.

¹¹⁸ That question was, however, later reopened, with the prediction of the existence of gravitational waves based on Einstein’s general theory of relativity. The experimental verification of gravitational waves has meanwhile given further strong support to a local understanding of gravitation. See Thorne (2017), and see Christensen and Meyer (2022) for a recent review article.

a difference to be found. The principle is indeed pretty sound in biology and also with respect to many natural phenomena outside biology—as the saying goes, even no two snowflakes are alike.¹¹⁹ But the principle has pretty silently been dropped from all considerations of really tiny entities, like elementary particles, atoms, or molecules. In that domain, there are significant effects due to the fact that there are fewer distinct ways of arranging a multitude of indiscernible (or indistinguishable) particles than there are ways of arranging the same number of discernible particles—the already mentioned phenomenon of Bose-Einstein condensation provides an example. The existence of indistinguishable particles is now taken as a rock bottom truth about the constitution of matter in the universe.¹²⁰ Quantum randomness is, in our view, presently at the interesting stage of solidifying into a similarly firmly accepted truth, which may mean that the deterministic itch discussed in Sect. 3.4 will, at least in that area, slowly fade away.

3.5.1 *Randomness for Technological Applications*

Even today, those who feel the deterministic itch strongly are selective about it. In some cases, randomness and indeterminism are simply wanted as a resource. Think of public lotteries, which are legal in most countries, with official rules that set standards of fairness. In a fair lottery, the initial state of the lottery machine is such that the numbers drawn are certified to be random, unless the machine fails, in which case the result is annulled and the procedure repeated. For a lottery machine, we actually require the full initial state to be such that it allows *only* for a list of possible outcomes and associated (usually uniform) probabilities. Coin-tossing is considered good enough a lottery machine in some cases (e.g., by the official rules, in determining the team to kick off a game of soccer). Coin-tossing has the great advantage of being simple to realise and being fully publicly observable and understandable, but it does not deliver very good randomness.¹²¹ The wheel of fortune familiar from school fairs is only slightly better. For official lotteries, much more complex mechanical lottery machines are common, but these can fail for many different reasons.¹²²

¹¹⁹ See Libbrecht (2021, 24–27) for an engaging discussion of this issue. Libbrecht concludes that “[e]ven with 10^{24} snow crystals being made per year on earth, the probability that any two would be exactly identical within the entire lifetime of the Universe is zero in any practical sense” (26).

¹²⁰ Philosophers still like to make fine distinctions, of course, and different grades of indistinguishability are still being discussed. See French (2019) for a well-informed overview of that debate.

¹²¹ See Diaconis et al. (2007) for a classic study of coin tossing. Trained magicians, such as Diaconis himself, can control the tossing such that they determine the side on which a fair coin lands to high accuracy. See Mackenzie (2009) for a popular account that also contains an image of a coin tossing machine that generates 100% heads with a standard, unbiased coin.

Lotteries may be highly special cases, and perhaps they point to remnants of superstition. Drawing lots to determine issues of some importance is an ancient practice that was often even tied to an epistemic reading of the outcome: the winning lot would indicate who was in fact favoured by the gods.¹²³ But in fact, our needs for reliable, secure randomness have increased immensely over the last few decades, far beyond lotteries and other long-established uses of randomness in games and in statistical sampling.¹²⁴ In science, the randomisation of experiments is now a standard requirement, especially in clinical trials in which one needs to exclude biases and selection effects, or in questionnaires, in which one randomises in order to counter question order effects. Assumptions about the free, randomised selection of experimental parameters are also crucially important, e.g., in interpreting quantum correlation experiments (see Footnote 80). Outside of experiments, high-quality randomness is furthermore needed for simulations (or computer experiments) and for a number of numerical so-called “Monte Carlo” techniques. But in terms of the amount of randomness, the biggest demand today is linked to security. Most modern communication meanwhile has some form of cryptographic protection, at the base of which is always some amount of randomness that is used to scramble messages. Even more randomness is used for securing computers against attacks, e.g., through so-called address space layout randomisation, which makes it harder for attackers to execute malicious code. Lots of randomness, that is, goes into protecting privacy and warding off malevolent manipulations.¹²⁵

In scientific and in security applications, traditional pseudorandomness (the algorithmic, deterministic computation of numbers that are hard to predict unless one knows the initial values) is increasingly found to be insufficiently reliable, and physical sources of randomness are therefore being used to improve the quality of the randomness.

¹²² For example, a notable malfunction in the German official Saturday lottery occurred on 11 May 2019. It was handled according to the rules that are in place for such incidents, and numbers were still drawn and published on that evening.—Some lotteries (e.g., the Swiss national lottery) have in fact meanwhile switched to using quantum-based randomness (see below) rather than mechanical devices.

¹²³ See Hacking (1990) for an account of how chance was associated with superstition until well into the nineteenth century.

¹²⁴ Again, Hacking (1990) provides useful background information.

¹²⁵ According to the U.S. National Institute of Standards, random numbers are used “hundreds of billions of times a day” (NIST, 2018). This means that our technological civilisation consumes randomness at a rate of around 10^{12} bits per day, or millions of bits per second, and the demand is rising fast.

3.5.2 *From General Physical to Quantum Random Generators*

Many modern computer processors offer dedicated sub-circuits through which locally sourced physical randomness, e.g., from thermal noise or from jitter in oscillators, is made available for applications, including the generation of secure cryptographic keys.¹²⁶ A number of physical randomness sources based on a variety of processes (collectively called “true random number generators”, or TRNGs) have been shown to pass all sorts of statistical tests for ascertaining randomness. The gold standard for such physical sources is, however, base-level quantum randomness, and there are vendors supplying products based on quantum randomness (i.e., on the basic truth of the Born rule) for use in smartphones, standard computers, and data centres.¹²⁷ Furthermore, there are government initiatives such as the U.S. NIST Randomness Beacon, which is meant to “promote the availability of trusted public randomness as a public utility”, to be “used for example for auditability and transparency of services that depend on randomized processes” (NIST, 2022). One use case here is the following: suppose I want to buy a thousand gadgets. The vendor guarantees that they all work. When the shipment comes I will want to examine a few to check, and I will want to have the right to refuse the lot if I find faulty ones. Which ones should I check? As a buyer, I will want to make sure the seller doesn’t nudge me to check gadgets from a select set that they have carefully checked themselves (while the rest may be junk). And as a seller, I want to make sure the buyer also acts fairly and samples really randomly. Given the NIST beacon, we can make a contract that ties the selection of which gadgets to check to random numbers that the beacon will publish in the future. None of us can know the outcome, as the randomness is certifiably quantum-based (Bierhorst et al., 2018). We enter a contract by which we both rely on the open future being settled one way or other. The NIST beacon is really a technology that incorporates the difference between lack-of-information “tell me more” uncertainty and open-future “wait and see” uncertainty in a very pure form: All details of the machinery are published, its full state at a given time may be known (no “tell-me-more” uncertainty), and yet nobody can know the random number that it will produce tomorrow at noon, simply because that future random number has not been produced yet. Like the outcome of any yet unperformed quantum experiment, that number simply does not yet exist. There are various possibilities for what that number may turn out to be, and these have (in this case, strictly uniform) probabilities attached, and that is all.

Technological developments have thus taken the philosophical conundrum of QM randomness to the next level, using features of the theory that were considered

¹²⁶ A critical discussion of the implementation of the Intel (thermal-noise based) RDRAND command is given by Ragab et al. (2021). See Wu et al. (2019) for a study of physical randomness from ring oscillators, which is, e.g., implemented on Apple processors.

¹²⁷ One such vendor is ID Quantique, based in Geneva (IDQ, 2022). See Herrero-Collantes and Garcia-Escartin (2017) for a review of technologies, and Mannalath et al. (2023) for exemplary tables listing products available in 2022.

conceptually puzzling to build technology for the whole world to rely on. Many quantum randomness generators are available for purchase as of 2024. The NIST beacon technology has been available for a number of years, with an update to version 2.0 in 2019, and development is ongoing. Further related technologies are more at the proof-of-principle stage, e.g., quantum cryptography and long-distance quantum communication. We cannot give an overview of these dynamically evolving fields here,¹²⁸ but we will mention one exciting development in the field of randomness, viz., device-independent randomness.

3.5.3 *Device-independent Randomness*

We said above that quantum randomness is the gold standard of physical randomness generation, because the Born rule is fundamental. Extending the metaphor, we could say that device-independent randomness is the gold standard of gold standards of randomness, because it produces randomness even independently of the formalism of QM, and in a way that is not based on any assumptions of the inner workings of the device that produces the randomness (hence the name). It is the intended technology of choice for the NIST randomness beacon mentioned above. A successful proof-of-principle experiment was reported by Pironio et al. (2010); see, e.g., Bierhorst et al. (2018) and Agresti et al. (2020) for recent advances. In the protocol of device-independent randomness, violations of Bell’s inequality are the crucial resource that provides the guarantee of randomness. This guarantee states that the resulting “random bits could not have been predicted according to any physical theory that prohibits faster-than-light (superluminal) signalling *and that allows independent measurement choices*” (Bierhorst et al., 2018, our emphasis). As long as you accept these highly plausible premises, you can be as sceptical as you wish about QM and still be assured of the fundamental randomness of the protocol’s output. As hinted at in the quote, however, agency scepticism would effectively undermine this result.

The general setting is as follows. Alice and Bob, both in securely shielded locations, want to generate random bits that cannot be predicted by a hypothetical adversarial eavesdropper, Eve, who has access to all information outside of their respective locations. If Alice and Bob can take recourse to a locally operating device for which they have a trusted model according to which Eve cannot know the device’s output, all will be well. But Alice and Bob may have been fooled by a vendor secretly cooperating with Eve, who sold them devices that output bits with impeccable statistics that, however, were pre-recorded and are known to Eve. It may be impractical for Alice and Bob to analyse their devices to such a level of detail that they can uncover the vendor’s fraud. What else can they do?

¹²⁸ For some relevant pointers to the literature, see, e.g., Ekert (1991) for a quantum cryptographic key distribution scheme directly based on Bell’s theorem, Zhang et al. (2022) for an advanced experimental realisation, and Briegel et al. (1998) for an entanglement-based scheme of a quantum repeater.

The protocol of device-independent randomness describes how Alice and Bob can devise a way of escaping their predicament if they have access to entangled quantum systems. In the set-up described by Bierhorst et al. (2018), they collect data from loophole-free Bell tests for which they control the individual settings at their locations—this is the assumption of agential freedom mentioned above. Furthermore they have to assume that superluminal signalling is impossible—but this is no more than assuming that their adversary, whatever her powers, is a physical entity.

Given their sequences of measurement setting selections and measurement outcomes, Alice and Bob can check whether they have witnessed violations of Bell’s inequality, indicating that they indeed had access to correlated quantum systems. In that case, they can use statistical methods of randomness extraction to generate high-quality random bits that Eve provably cannot predict. We will not go on to describe the concrete mathematics, for which see Bierhorst et al. (2018), but at bottom, the method comes down to the fact that for Alice, knowing a state of a correlated joint system such as the Bell singlet state $|\psi_s\rangle$ of Eq. (3.102) entails that her prediction of her own measurement outcome is fundamentally probabilistic, as expressed in Eq. (3.108).

Device-independent randomness is a quantum technology because it harnesses the kind of correlations that come about via quantum entanglement of spatially separated parts of a joint system. But any physical system that produces correlations strong enough to violate Bell’s inequality could be plugged into the protocol. In principle, thus, the protocol is not necessarily dependent on entangled particles—it is just that as far as we know, these are the only systems in nature that produce correlations of the required strength.

3.5.4 *Noise and Quantum Error Correction*

In this section, we will introduce some basic ideas of quantum error correction, which is an essential protocol for any large-scale applications in which quantum information is stored and processed in complex quantum devices. The successful execution of quantum error correction will require automated procedures that operate with high fidelity. Arguably, it will also require procedures with a certain level of flexibility and adaptivity to deal with variable types of imperfections and noise in such devices. The material that we will discuss here, together with the material of Sect. 3.4.4, will also provide some background for Chap. 5 where we will study artificial learning agents that can be successfully employed for quantum error correction, and other tasks in quantum environments.

Before we start with quantum error correction, however, we will first elaborate on the description of noise. Noise is modelled as an irreversible process in quantum physics, and we will link to our discussion of that topic in earlier sections.

3.5.4.1 The Description of Noise

In the discussion of the Born rule and when to use it (Sect. 3.3.2.2) we briefly reconsidered nuclear fusion processes in the p-p-chain, which are irreversible inelastic scattering processes described by quantum field theory. The Born rule is used in the computation of the cross section, or rate of occurrence, of these scattering processes. The probability that an incoming (“initial”) state $|i\rangle$ scatters into an outgoing (“final”) state $|f\rangle$ under the influence of a unitary scattering operator U is given by

$$\Pr(i \rightarrow f) = |\langle f|U|i\rangle|^2. \quad (3.143)$$

For the initial and final states appearing in Eq. (3.143) one typically assumes particles of definite momenta, that is, in the Heisenberg picture, eigenstates of the momentum observables at different times ($t \rightarrow -\infty$; “long before” and $t \rightarrow +\infty$; “long after” the interaction), as illustrated in Fig. 3.7.

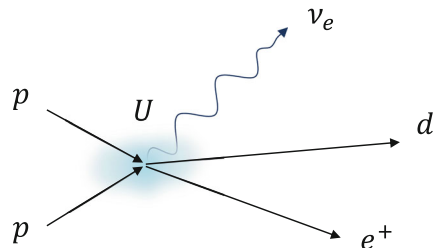
The lowest-order contribution to Eq. (3.143) is obtained by a perturbation expansion of the scattering operator in the Hamiltonian that describes the interaction of the quantum fields associated with the particles. For the first reaction in the p-p-chain (Eq. (3.62)), one thereby obtains the overall rate for proton-proton capture and conversion into deuterium (Bahcall & May, 1969)

$$R_{pp} = \frac{2\pi}{\hbar V} \sum_{\text{spins}} \int d^3n_{p_1} d^3n_{p_2} \Phi(E_1) \Phi(E_2) d^3n_d d^3n_e d^3n_\nu \delta(E_i - E_f) |\langle f|H_\beta|i\rangle|^2, \quad (3.144)$$

where H_β is the Hamiltonian that generates the scattering transitions (which also describes beta decay), $d^3n_i = V/(2\pi\hbar)^3 d^3p_i$, and $\Phi(E_i)$ are the Maxwell-Boltzmann distributions for the two incoming protons. The integral describes the overall rate (proportional to the scattering probability) with which two protons are scattered into deuterium and the other reaction products.¹²⁹

Note that the integration over the outgoing momenta, i.e., the momenta p_d , p_e , p_ν , as well as the summation over the spin degrees of freedom of all particles is

Fig. 3.7 Schematic illustration of the inelastic scattering process describing the first step of the p-p-chain corresponding to Eq. (3.62)



¹²⁹ Further details can, of course, be added; see for example Carlson et al. (1991). For our overall argument, the formalities of the calculation are not decisive.

performed if one is only interested in the particle conversion rate. The integration over the momenta p_1, p_2 of the two incoming protons, which are sampled from a Maxwell-Boltzmann distribution, takes account of the situation that in the sun all of these configurations will occur with probabilities given by the thermal distribution. Even if one were to try and keep track of the motional state of all protons (and maybe the other nuclei and particles in the solar plasma and even beyond), tracing out a small subset of these particles would quickly lead to a mixed state that is diagonal in the momentum basis.¹³⁰ So one is perfectly warranted to describe the incoming state of the two protons, that is, the source of the scattering “experiment” in the terminology of Sect. 3.3.2.1, as a (mixed) state of definite momenta.

Noise in Quantum Information The objective of the field of quantum information is to exploit the properties of elementary quantum systems, such as quantum coherence and entanglement, controlling them with sufficient precision to enable novel applications. We have given a few examples in Sect. 3.4.4 above, including device-independent randomness generation, dense coding, quantum teleportation, and one-way quantum computation, but there are many more applications in quantum communication and quantum computing that are currently being explored. For most of these applications it is essential that the systems are prepared in an almost perfectly pure state and maintained with a high degree of quantum coherence over sufficiently long times. This typically requires almost perfect isolation of the quantum systems whose states carry the quantum information against unwanted interactions with the environmental degrees of freedom.¹³¹ No agent can keep track of, let alone control, these irreversible interactions.

Consider a single qubit, initially prepared in some state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$. The effect of noise and uncontrolled interactions with its environmental degrees of freedom can be described, quite generally, as follows. If we include the environment into the description, corresponding to a (hypothetical) situation in which we have full information about the quantum state of its degrees of freedom, then the initial state of the total system can be written in the form $|i\rangle = |\psi\rangle|e\rangle$. The probability that, at a later time, the system and environment would be found in some other state $|f\rangle$ is then given as $\Pr(i \rightarrow f) = |\langle f|U|i\rangle|^2$, similar as in Eq. (3.143), where U describes the dynamical evolution of the total system (Fig. 3.8).¹³²

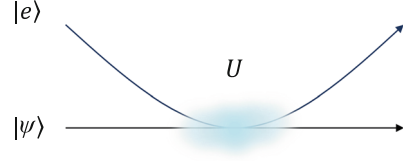
Let us first consider the case where the qubit is initially in one of its computational basis states, say $|0\rangle$. The corresponding transformation then reads as follows:

¹³⁰ In fact this would already be the case for the much simpler situation of a non-relativistic gas of N particles that only interact via elastic collisions. We will come back to a similar situation in the next paragraph.

¹³¹ In an ion-trap quantum computer, for example, fluctuations of ambient magnetic fields of the ions lead to dephasing (that is, a loss of quantum coherence) of superpositions of the ions’ internal states with different magnetic moments (Häffner et al., 2008; Schindler et al., 2013).

¹³² Note that, in the Heisenberg picture, U transforms between the eigenstates of observables for the combined system consisting of the qubit and the environment at different times.

Fig. 3.8 Schematic illustration of the interaction of a single qubit, prepared in state $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, with its environment, which is initially in some state $|e\rangle$



$$\begin{aligned} |0\rangle|e\rangle &\longrightarrow |0\rangle|e_{00}\rangle + |1\rangle|e_{01}\rangle, \\ |1\rangle|e\rangle &\longrightarrow |1\rangle|e_{11}\rangle + |0\rangle|e_{10}\rangle, \end{aligned} \quad (3.145)$$

where the $|e_{ss'}\rangle$ ($s, s' \in \{0, 1\}$) denote the (unnormalized) final states of the environment corresponding to the transitions $|s\rangle \rightarrow |s'\rangle$ of states of the qubit. Equation (3.145) is quite general and simply expresses the fact that, at any later time, the total state of the system and the environment, when expanded in the computational basis of the qubit, will have at most two terms.¹³³ By linearity of U one obtains

$$(\alpha|0\rangle + \beta|1\rangle)|e\rangle \longrightarrow \alpha(|0\rangle|e_{00}\rangle + |1\rangle|e_{01}\rangle) + \beta(|1\rangle|e_{11}\rangle + |0\rangle|e_{10}\rangle) \equiv |\Psi'\rangle. \quad (3.146)$$

By rearranging the different terms, similar as in the derivation of the teleportation protocol (Eqs. (3.127) and (3.128)), the scattered state $|\Psi'\rangle$, i.e., the joint state of the qubit and the environment after the interaction, can be rewritten in the form

$$|\Psi'\rangle = \sum_{\mu=0}^3 (\sigma_{\mu}|\psi\rangle) |e_{\mu}\rangle, \quad (3.147)$$

where $|e_{\mu}\rangle$ are linear combinations of the environmental states $|e_{ss'}\rangle$ that appear in Eq. (3.145).¹³⁴ From Eq. (3.147) one can derive a general expression for the reduced state of the qubit alone, by taking the partial trace over the environmental degrees of freedom, to obtain

$$\rho' = \text{tr}_{\text{env}}(|\Psi'\rangle\langle\Psi'|) = \sum_{\mu, \nu=0}^3 \sigma_{\mu}|\psi\rangle\langle\psi|\sigma_{\nu}\langle e_{\nu}|e_{\mu}\rangle. \quad (3.148)$$

The terms $\langle e_{\nu}|e_{\mu}\rangle$ that appear in the summation in Eq. (3.147) depend on the concrete physical realisation of the environment. To give an example, let us assume that the environment consists of a single, uncontrolled spin which interacts with the system qubit via a quantum Ising interaction described by the Hamilton operator

¹³³ For weak interactions we have, in addition, $|e_{00}\rangle \simeq |e\rangle \simeq |e_{11}\rangle$, from which it follows that $\langle e_{00}|e_{00}\rangle, \langle e_{11}|e_{11}\rangle \simeq 1$ and $\langle e_{01}|e_{01}\rangle, \langle e_{10}|e_{10}\rangle \ll 1$.

¹³⁴ Specifically, one obtains $|e_0\rangle = (|e_{00}\rangle + |e_{11}\rangle)/2$, $|e_1\rangle = (|e_{01}\rangle + |e_{10}\rangle)/2$, $|e_2\rangle = (|e_{01}\rangle - |e_{10}\rangle)/(2i)$, $|e_3\rangle = (|e_{00}\rangle - |e_{11}\rangle)/2$.

$$H = \hbar g(t) \sigma_z^{(a)} \sigma_z^{(b)}, \quad (3.149)$$

where the interaction strength $g(t)$ has a stochastic time dependence, and the indices a and b refer to the system and the environmental qubit, respectively. For a qubit realized in the internal states of a trapped atom, for example, a physical realization could consist in internal-state-dependent dephasing collisions between the system qubit and other atoms from the environment (e.g., from the residual background gas in an experiment with an imperfect vacuum).¹³⁵ The unitary scattering operator for such a collision is obtained by integrating the dynamics with the Hamiltonian of Eq. (3.149),

$$U(\varphi) = e^{-i\varphi \sigma_z^{(a)} \sigma_z^{(b)}}, \quad (3.150)$$

where $\varphi = \int dt g(t)$ is a random phase factor. The precise probability distribution $w(\varphi)$ for the accumulated phase φ does not matter for the following considerations, but it will depend on the details of the collision process, e.g., the temperature (determining the thermal velocity distribution of the gas) and the atomic interaction strength. In the following we assume the environmental spin to be in a random initial state $|e\rangle = a|0\rangle + b|1\rangle$, with $|a|^2 + |b|^2 = 1$.¹³⁶ If we average Eq. (3.148) over random initial states of the environmental spin and over the distribution $w(\varphi)$ of scattering phases, we obtain¹³⁷

$$\rho' = (1 - p) \rho + p \sigma_z \rho \sigma_z = (1 - p) \rho + p Z \rho Z, \quad (3.151)$$

where $\rho = |\psi\rangle\langle\psi|$ denotes the initial state of the system qubit and $p = \langle \sin^2(\varphi) \rangle$, using “ $\langle \dots \rangle$ ” to indicate the average over the random phase factor. Equation (3.151) is an example of a quantum noise channel that in the context of quantum error correction is known as the *dephasing channel*, or the *phase-flip channel*.

Similar considerations can be carried out for other types of interactions of the qubit with the environment. For example, collisions that can change the internal energy states of the qubit can be modelled by the Hamiltonian $H = \hbar g(t) \sigma_x^{(a)} \sigma_x^{(b)}$.

¹³⁵ The main sources of dephasing for qubits realized in trapped atoms are fluctuations of the magnetic field and the electromagnetic trapping potential (Saffman, 2016), while uncontrolled collisions with residual particles seem to play only a minor role. For a simple illustration of our main argument, however, it is instructive to consider a situation with a spin-type environmental degree of freedom.

¹³⁶ By this we mean a flat distribution of the initial states $|e\rangle = a|0\rangle + b|1\rangle$ under the Haar measure, which can be visualized as equally distributed endpoints of the unit vector $\langle \vec{\sigma} \rangle = \text{tr}(\rho \vec{\sigma})$, associated with the state $\rho = |e\rangle\langle e|$, over the surface of the two-dimensional (Bloch) sphere S^2 (Nielsen & Chuang, 2010, Ch. 1.2). For this manifold of states, $|a|^2 > |b|^2$ ($|a|^2 < |b|^2$) corresponds to states on the northern (southern) hemisphere of S^2 .

¹³⁷ The main steps are to show that $\langle e_0|e_0 \rangle = \sin^2(\varphi)$, $\langle e_3|e_3 \rangle = \cos^2(\varphi)$, $\langle e_3|e_0 \rangle = i(|a|^2 - |b|^2) \sin(\varphi) \cos(\varphi) = -\langle e_0|e_3 \rangle$, and $\langle e_\nu|e_\mu \rangle = 0$ for all other values of ν and μ . This one obtains by first deriving an explicit expression of the states $|e_{ss'}\rangle$ in Eq. (3.146) for $U = U(\varphi)$ and then employing the transformation as in Footnote 134.

The effect of such a collision (with otherwise identical assumptions) on the qubit's reduced density matrix is described by

$$\rho' = (1 - p) \rho + p X \rho X, \quad (3.152)$$

which in quantum error correction is also known as the *bit-flip channel*.

Quantum vs. Classical Noise To understand the “bit flip” and “phase flip” terminology for quantum noise channels, consider the effect of different Pauli operators acting on the initial state of the qubit. When classical information is sent through a noisy communication channel, classical bit flips are transitions of the form $0 \leftrightarrow 1$, which can alter a message in a random way. The quantum analogue of a bit flip in the Hilbert space of a qubit is described by the Pauli σ_x operator, which acts on the basis states as

$$X|0\rangle = |1\rangle, \quad X|1\rangle = |0\rangle. \quad (3.153)$$

While bit flips are the only errors that can happen in the classical case, the situation is different for quantum information. Since a quantum bit is specified by the complex amplitudes α and β in $|\psi\rangle = \alpha|0\rangle + \beta|1\rangle$, any change of these amplitudes can alter the message. An example of such a change is a *phase flip*, which has no classical analogue and is described by the transition

$$|\psi\rangle = \alpha|0\rangle + \beta|1\rangle \longrightarrow \alpha|0\rangle - \beta|1\rangle = Z|\psi\rangle. \quad (3.154)$$

Note that in the resulting state $Z|\psi\rangle$, the probabilities $|\alpha|^2$ and $|\beta|^2$ to detect, respectively, the bit values 0 or 1 via a σ_z measurement are the same as in the state $|\psi\rangle$. However, the altered value $\beta \rightarrow -\beta$ in Eq. (3.154) makes a difference if the qubit is measured in a different basis, i.e., if an observable different from σ_z is measured. This option does not exist for classical information.

Using the above terminology, we can now interpret the mixed state in Eq. (3.151) as follows. As a result of dephasing collisions (as described in Fig. 3.8, Eqs. (3.146) and (3.150)), a phase flip happens with probability $p = \langle \sin^2(\varphi) \rangle$, resulting in the state $\rho = Z|\psi\rangle\langle\psi|Z$, while with probability $1 - p = \langle \cos^2(\varphi) \rangle$ the qubit remains in the initial state $\rho = |\psi\rangle\langle\psi|$.¹³⁸ This description of the qubit is warranted when the interactions with the environment are irreversible. In the current

¹³⁸ Using a similar analysis, one can show that additional collisions of the system qubit with other environmental particles will maintain the form of Eq. (3.151), but with averages over multiple random phase factors appearing in the mixture. Furthermore, each of the environmental qubits will itself be involved, over a time scale related to the mean free path length (or the collision rate) of a Boltzmann gas, in similar dephasing collisions with other environmental qubits from the gas. As a result, the irreversibility of the collisional interactions warrants a state assignment where each of the qubits (for any chosen subgroup of qubits in the gas) is completely dephased. In Calsamiglia et al. (2005) this scenario was studied for the special case where all qubits of the gas are initially spin-polarized in the x -direction, corresponding to the choices $|a|^2 = 1/2 = |b|^2$ in Footnote 136.

example we considered such a situation, assuming dephasing collisions with a gas of environmental spins similar as—even though less dramatic than—the situation with nuclear collisions in the sun. There is nothing any agent could do to reverse such evolutions. Something has happened that cannot be undone.

Of course there are also situations when the interactions between many qubits *are* reversible. One can in fact realize a situation where qubits do not move around and collide randomly, as in a gas, but where their motion is controlled and each of the qubits is trapped on a specific site of a lattice, for example, in an optical lattice (Jaksch et al., 1999). Under such conditions, highly controlled “ultra-cold” collisions, described by a Hamiltonian similar to Eq. (3.149), can be implemented such that one can control and tune the interaction phase φ in Eq. (3.150) with high precision. Such agent-controlled interactions can then be used to transform the entire lattice of qubits into highly entangled states, such as the universal cluster state introduced by Briegel and Raussendorf (2001). These experimental situations can be realized in modern-day quantum optics laboratories.¹³⁹ To describe them, one is *not* warranted to assign to each of the qubits on the lattice a mixed state, because this would not account for the experimentally observable properties of larger groups of qubits in the lattice. The capacity of each qubit to interfere with its neighbours is instead a key feature that is exploited in such experiments. The spin interactions in such experiments are fully reversible.

General Description of Noise In a matrix representation, the action of the dephasing channel on a qubit that is initially in the pure state $\rho = |\psi\rangle\langle\psi|$ has the following form:

$$\begin{pmatrix} |\alpha|^2 & \alpha\beta^* \\ \alpha^*\beta & |\beta|^2 \end{pmatrix} \longrightarrow \begin{pmatrix} |\alpha|^2 & (1-2p)\alpha\beta^* \\ (1-2p)\alpha^*\beta & |\beta|^2 \end{pmatrix}. \quad (3.155)$$

Here one can see that, in the limit $p \rightarrow 1/2$, the off-diagonal entries (“coherences”) vanish, which corresponds to a loss of the qubit’s capacity for interference. The same factor $1 - 2p$ that multiplies and dampens the off-diagonal matrix elements in Eq. (3.155) also appears in the discussion of the duality relation (see Footnote 71).

A more general noise channel that is often used in quantum information is the so-called *Pauli channel*:

$$\rho \longrightarrow \rho' = \sum_{\mu=0}^3 p_{\mu} \sigma_{\mu}^{(a)} \rho \sigma_{\mu}^{(a)} = p_0 \rho + p_1 X_a \rho X_a + p_2 Y_a \rho Y_a + p_3 Z_a \rho Z_a. \quad (3.156)$$

It describes a noise process, acting on qubit a , where one of four transitions (also called “physical errors”) happen: no error, a bit flip, a bit-phase flip, or a phase flip, described by operators I_a , X_a , Y_a , or Z_a , respectively, with probabilities p_0 , p_1 , p_2 , and p_3 , where $p_0 = 1 - p_1 - p_2 - p_3$. For the special choices $p_1 = p$, $p_2 = 0 = p_3$

¹³⁹ For a recent reference see, e.g., Bluvstein et al. (2022).

one obtains the previously described bit-flip channel (Eq. (3.152)), while $p_1 = 0 = p_2$, $p_3 = p$ yields the phase-flip channel (Eq. (3.151)). For the symmetric choices $p_1 = p_2 = p_3 = p/4$, the resulting channel is referred to as the *depolarizing channel*.¹⁴⁰

The most general description of a noise process on a quantum system that is initially in a pure state is given by a linear, completely positive and trace preserving map (Lindblad, 1976; Alicki & Lendi, 2007).¹⁴¹ Complete positivity maintains the positivity of a density operator, including its trivial extension to a larger Hilbert space.¹⁴² Every completely positive and trace preserving map on a density operator ρ can be represented in the form (Nielsen & Chuang, 2010, Ch. 8.2)

$$\rho \longrightarrow \rho' = \sum_k E_k \rho E_k^\dagger, \quad (3.157)$$

where $\{E_k\}$ is a set of linear, so-called Kraus operators satisfying the relation

$$\sum_k E_k^\dagger E_k = \hat{1}. \quad (3.158)$$

If the quantum system is a single qubit, one easily recovers the Pauli channel (Eq. (3.156)) for the choice $E_k = \sqrt{p_k} \sigma_k$, $k = 0, \dots, 3$, but other choices are possible where the E_k are neither Hermitian nor unitary operators (in contrast to the Pauli operators, which are both), for example to model the process of *spontaneous decay*.

A completely positive map as given in Eq. (3.157) can also be used for a more general description of a quantum measurement which includes noisy or otherwise imperfect equipment, in which case the map is no longer trace preserving. Any quantum measurement can thus be described by a set of linear operators $\{E_k\}$ as in Eq. (3.158), which specifies a corresponding set of measurement outcomes $\{k\}$, probabilities, and state updates through the relations

$$\text{Pr}(k) = \text{tr}(E_k^\dagger E_k \rho) \quad (3.159)$$

¹⁴⁰ The suitability of these models depends on the noise that needs to be taken into account, which in turn depends on the details of the interactions of the physical qubits with the environment. More details about the treatment of such physical noise processes in the context of quantum error correction will be given in the next section.

¹⁴¹ The same holds for any quantum system, as long as the joint state of the system and the environment are initially disentangled. In general, however, it need not be true (Pechukas, 1994).

¹⁴² Positivity of the completely positive and trace preserving map implies that it maps any density operator (which, per definition, has only eigenvalues that are positive or zero) to a new density operator. *Complete* positivity means that this property is maintained for any extended map that acts like the identity outside the Hilbert space on which the density operator is defined.

and

$$\rho \rightarrow \frac{E_k \rho E_k^\dagger}{\text{tr}(E_k^\dagger E_k \rho)}. \quad (3.160)$$

These relations generalise the corresponding expressions in the context of the Born rule, Eqs. (3.45) and (3.48), respectively. If we write $\hat{O}_k = E_k^\dagger E_k$ then Eqs. (3.158) and (3.159) describe a generalised measurement, or POVM, as introduced in Eq. (3.51). As we noted already in Sect. 3.3.2.2, a POVM does not specify how the quantum state should be updated. However, specific choices of Kraus operators of the completely positive map described by Eq. (3.160) allow us to model different scenarios. For ideal measurements described by orthogonal projectors, we have $E_k = \hat{P}_k = E_k^\dagger$,¹⁴³ and thus Eq. (3.160) reduces to Eq. (3.48). However, one can also take E_k to be a linear combination of different projectors. Consider, for example, the measurement of the observable $\hat{S}_z = (\hbar/2)\sigma_z$ of a single spin-1/2 particle, or qubit, as in our discussion of the Stern-Gerlach experiment. In the ideal case, the corresponding projectors are $\hat{P}_0 = \hat{P}_{|\uparrow, z\rangle} = |\uparrow, z\rangle\langle\uparrow, z|$ and $\hat{P}_1 = \hat{P}_{|\downarrow, z\rangle} = |\downarrow, z\rangle\langle\downarrow, z|$. Now consider the combinations

$$\begin{aligned} E_0 &= \sqrt{\eta} \hat{P}_0 + \sqrt{1-\eta} \hat{P}_1, \\ E_1 &= \sqrt{\eta} \hat{P}_1 + \sqrt{1-\eta} \hat{P}_0 \end{aligned} \quad (3.161)$$

with $0 \leq \eta \leq 1$. For the spin (qubit) prepared in the pure state $\rho = |0\rangle\langle 0| = |\uparrow, z\rangle\langle\uparrow, z|$, we thus obtain

$$\begin{aligned} \Pr(k=0) &= \text{tr}(E_0^\dagger E_0 |0\rangle\langle 0|) = \text{tr}(\hat{O}_0 |0\rangle\langle 0|) = \eta, \\ \Pr(k=1) &= \text{tr}(E_1^\dagger E_1 |0\rangle\langle 0|) = \text{tr}(\hat{O}_1 |0\rangle\langle 0|) = 1 - \eta \end{aligned} \quad (3.162)$$

for the probabilities $\Pr(0)$ and $\Pr(1)$ of detecting the outcomes “0” and “1”, respectively. Similar relations are obtained for the choice $\rho = |1\rangle\langle 1|$. These relations describe an imperfect measurement process, where the measurement outcome is correct only with a certain probability η , while it is wrong with probability $1 - \eta$. For example, even though the system was prepared in the pure state $|0\rangle$ (which means that the measurement of σ_z , with an ideal Stern-Gerlach apparatus described by $\{\hat{P}_0, \hat{P}_1\}$, would give the outcome “0” with certainty), the measurement with the imperfect apparatus described by the POVM $\{\hat{O}_0, \hat{O}_1\}$ would report the wrong result “1” in $1 - \eta$ of all cases.

In the next section we will leave our brief introduction to the description of noise channels and generalised measurements behind and discuss methods to protect quantum information that is encoded in the physical state of a system from noise and from the unwanted but inevitable interactions of that system with the environment.

¹⁴³ Note that in this case we also have $E_k^\dagger = \hat{P}_k^\dagger = \hat{P}_k$ and thus $E_k^\dagger E_k = \hat{P}_k^\dagger \hat{P}_k = (\hat{P}_k)^2 = \hat{P}_k$.

3.5.4.2 Quantum Coding and Error Correction

Different methods have been developed to protect the information represented by α and β in Eq. (3.77) and to undo the effect of noise on a single spin-1/2 system representing the qubit (Terhal, 2015; Lidar & Brun, 2013). The basic idea is to delocalise the information over several physical systems, or degrees of freedom, and to use methods inspired by redundant coding.

In classical communication, a simple method is to send an information bit in multiple copies (e.g., 000 instead of 0) through a noisy communication channel, that is, to use a simple *repetition code*: $i \rightarrow iii$ for $i \in \{0, 1\}$. The number of physical bits in the code (in this example, three) depends on the expected level of noise in the channel, i.e., the probability with which the value of a specific bit flips during transmission. If this probability is small, the receiver can simply measure the values of all bits and use a majority rule to infer the correct bit value with high probability.

In the quantum case, the protection of quantum information needs to be implemented in a fundamentally different way: the idea of the repetition code cannot be directly used, due to the no-cloning theorem (Dieks, 1982; Wootters & Zurek, 1982). According to this theorem, quantum information cannot be copied: a hypothetical process $|\psi\rangle \mapsto |\psi\rangle \otimes |\psi\rangle \otimes |\psi\rangle$ that maps arbitrary quantum states $|\psi\rangle \in \mathcal{H}$ to multiple copies, in analogy to $i \rightarrow iii$ for $i \in \{0, 1\}$ in the classical case, cannot be physically realised as it would violate the linearity of quantum mechanics and even allow faster-than-light communication (Herbert, 1982). However, an encoding that is reminiscent of the above repetition code is given by the following unitary mapping:¹⁴⁴

$$\begin{aligned} |\psi\rangle|00\rangle &= (\alpha|0\rangle + \beta|1\rangle)|00\rangle \rightarrow \alpha|000\rangle + \beta|111\rangle \\ &= \alpha|0\rangle_L + \beta|1\rangle_L \\ &= |\psi\rangle_L \end{aligned} \tag{3.163}$$

where the quantum codewords $|000\rangle \equiv |0\rangle_L$ and $|111\rangle \equiv |1\rangle_L$ are the analogues of the classical codewords 000 and 111, respectively. The quantum state $|\psi\rangle_L$ on the right hand side of Eq. (3.163) describes the *logical qubit*, with the same amplitudes α and β as in Eq. (3.77) but now encoded in three physical qubits. In the following, we will refer to Eq. (3.163) as a *quantum repetition code*. This is a simple code that could be used for a quantum memory.

¹⁴⁴ This encoding maps the Hilbert space of a single physical qubit into a two-dimensional subspace of the 2^n -dimensional Hilbert space of n physical qubits, which is called the code space (in the main text, we chose $n = 3$). Often the term *quantum code* is used for the code space, but sometimes also for the mapping that maps the qubit into the code space. In any case, it is sufficient to specify the quantum codewords $|0\rangle_L$ and $|1\rangle_L$, which are the images of the basis states $|0\rangle$ and $|1\rangle$ of the physical qubit and span the code space.

The information encoded in $|\psi\rangle_L$ is protected against a bit flip error on any one of the physical qubits. Consider, for example, a bit flip on the second physical qubit, which changes the quantum code words $|000\rangle$ and $|111\rangle$ into $|010\rangle$ and $|101\rangle$, respectively. To detect that error and undo its effect without destroying the quantum superposition, one cannot simply measure and compare the values of the individual physical qubits, as one would in the classical case. This would project the logical qubit $|\psi\rangle_L$ into one of the product states, e.g., $|010\rangle$ or $|101\rangle$, and thus destroy the information contained in the quantum superposition. However, one can measure certain *joint observables* M_k ($k = 1, 2$) of the physical qubits, whose eigenvalues 0 or 1 only carry relative (parity) information, indicating whether the possible values of neighbouring bits in the quantum superposition state are still the same, or different, respectively. In the example given above, these parity observables would be $M_1 = Z_1 Z_2$ and $M_2 = Z_2 Z_3$. A measurement of these observables does not destroy the logical information contained in the amplitudes α and β , but it does allow an agent to identify uniquely on which of the qubits a bit flip has happened.

To see this, we note that each of the code words, $|000\rangle$ and $|111\rangle$, and thus any superposition state $|\psi\rangle_L = \alpha|0\rangle_L + \beta|1\rangle_L$ of the logical qubit, is an eigenstate of both M_1 and M_2 , with the eigenvalue $+1$ in each case. In contrast, if a bit flip occurs on the first physical qubit, the resulting state $X_1|\psi\rangle_L$, together with the flipped basis states $|100\rangle$ and $|011\rangle$, is then an eigenstate of M_1 and M_2 with corresponding eigenvalues -1 and $+1$, respectively. Similar results apply for a bit flip X_j occurring on one of the other physical qubits $j = 2, 3$. The different cases of no bit-flip error or a bit-flip error on qubit 1, 2, or 3, thus correspond to the follow pairs of eigenvalues, or *error syndromes*, $(+, +)$, $(-, +)$, $(-, -)$, and $(+, -)$, respectively. By measuring the two joint observables M_1 and M_2 of the quantum register of three physical qubits, one can thus detect and identify any single bit-flip error. Furthermore, any such error can be undone by applying one of the following unitary correction operations $U_j = (X_j)^{-1} = X_j$ for $j = 1, 2, 3$, which rotate the logical qubit back into its original state:

$$U_j X_j |\psi\rangle_L = |\psi\rangle_L. \quad (3.164)$$

The encoding in Eq. (3.163) can thus protect the logical qubit against single bit-flip errors on any of the physical qubits. Consider the probability of a bit flip occurring on a specific physical qubit to be p . Without encoding, the quantum information that was carried on the single physical qubit would thus be changed with probability $p_L = p$. With the encoding of Eq. (3.163), the situation is different. The probability that a bit flip occurs on more than just one of the three physical qubits—a situation which cannot be corrected by that code—is given by $p_L = 1 - (1 - p)^3 - 3p(1 - p)^2$. One can easily check that, if $p < 1/2$, the encoding reduces the probability that the error remains undetected, $p_L < p$. For sufficiently small p , we have $p_L \simeq 3p^2$ which means that the probability of a *logical error*, is reduced quadratically. A problem with the repetition code of Eq. (3.163) is that it cannot detect phase-flip errors, described by error operators Z_j which appear, for example, in the dephasing channel (Eq. (3.151)) or in the Pauli channel

(Eq. (3.156)). Phase-flip errors do not affect the code words $|000\rangle$ and $|111\rangle$, apart from a possible phase factor in front of $|111\rangle$. In other words, these errors can only change the phase of the amplitude β (which is then mapped to $-\beta$), thereby mapping the logical qubit to a different logical qubit. The states $|\psi\rangle_L$ and $Z_j|\psi\rangle_L$ are thus in the same code space $\mathcal{H}_L = \text{span}\{|000\rangle, |111\rangle\}$ and an error cannot be detected.

3.5.4.3 Stabilizer Codes

The code in Eq. (3.163) served to illustrate the basic ideas of quantum error correction, exploiting its similarity with the classical repetition code. The quantum repetition code is an example of a more general class of quantum codes called *stabilizer codes* (Gottesman, 1997; Nielsen & Chuang, 2010). Similar as with stabilizer states, introduced in Eq. (3.130), a stabilizer code is defined as the common *eigenspace* with eigenvalue $+1$ of a set of commuting multi-qubit Pauli observables M_k , defined in Eq. (3.131). More specifically, a stabilizer code that uses N physical qubits to encode one logical qubit is defined as the common eigenspace of $N - 1$ of such independent observables $M_1, M_2, \dots, M_{N-1}$.

An example of a stabilizer code that protects a single qubit from general one-qubit errors is the Shor code, which uses $N = 9$ qubits and has the generators

$$\begin{aligned} M_1 &= Z_1 Z_2, & M_2 &= Z_2 Z_3, & M_3 &= Z_4 Z_5, \\ M_4 &= Z_5 Z_6, & M_5 &= Z_7 Z_8, & M_6 &= Z_8 Z_9, \\ M_7 &= X_1 X_2 X_3 X_4 X_5 X_6, & M_8 &= X_4 X_5 X_6 X_7 X_8 X_9. \end{aligned} \quad (3.165)$$

The code words of the Shor code have the form

$$\begin{aligned} |0\rangle_L &= \frac{1}{2\sqrt{2}} (|000\rangle + |111\rangle) (|000\rangle + |111\rangle) (|000\rangle + |111\rangle), \\ |1\rangle_L &= \frac{1}{2\sqrt{2}} (|000\rangle - |111\rangle) (|000\rangle - |111\rangle) (|000\rangle - |111\rangle), \end{aligned} \quad (3.166)$$

which are built from (tensor products of) GHZ-Mermin-type entangled states (Greenberger et al., 1990; Mermin, 1990). The latter correspond—up to a local change of basis—to graph states of three qubits, as displayed in Fig. 3.6c for $N = 3$. Note that the first six observables $M_1, \dots, M_6$, are identical to stabilizer generators of the quantum repetition code, comparing the bit values of neighboring bits in the groups $\{1, 2, 3\}$, $\{4, 5, 6\}$, and $\{7, 8, 9\}$. However, in addition there are two operators of a different form, M_7 and M_8 , whose eigenvalues ± 1 tell the agent whether the phases of the entangled states of two neighboring groups are the same or different. The code space $\mathcal{H}_L = \text{span}\{|0\rangle_L, |1\rangle_L\}$, spanned by the two codewords in Eq. (3.166), is the common eigenspace with eigenvalue $+1$ of all generators $M_1, \dots, M_8$ in Eq. (3.165):

$$M_k \mathcal{H}_L = \mathcal{H}_L, \quad k = 1, 2, \dots, 8, \quad (3.167)$$

thus any superposition of the form

$$|\psi\rangle_L = \alpha|0\rangle_L + \beta|1\rangle_L \in \mathcal{H}_L \quad (3.168)$$

represents an encoded qubit.

Stabilizer codes are similar to stabilizer states, the difference lies in the fact that the eigenvalues of the stabilizer generators do not determine a single state, but a two- or higher-dimensional subspace of the Hilbert space of N qubits. A crucial property of the stabilizer code is how the error operators that appear in a given noise channel, such as the bit-flip operator (X) and the phase flip operator (Z), which act on the physical qubits, commute with the generators of the stabilizer. For example, consider a bit-flip on the second qubit, described by the error operator X_2 . It is easy to check that

$$\begin{aligned} M_1 X_2 &= -X_2 M_1, \\ M_2 X_2 &= -X_2 M_2, \\ M_k X_2 &= X_2 M_k, \quad k = 3, \dots, 8. \end{aligned} \quad (3.169)$$

By measuring the eigenvalues of all stabilizers, which constitute the syndrome, one can detect the error. A bit flip X_2 on qubit 2 thus leads to an error syndrome of the form $(-, -, +, +, +, +, +, +)$, which is to be compared with $(+, +, +, +, +, +, +, +)$ in the case without a bit flip. Similar relations can be found for a bit- or phase-flip on any of the nine physical qubits. Importantly, any error operator $\sigma_\mu^{(a)}$ as in Eq. (3.156), applied to any of physical qubits $a = 1, \dots, 9$ of the Shor code, rotates the code space $\mathcal{H}_L$ into an *error space* $\sigma_\mu^{(a)} \mathcal{H}_L$, which is just a name for another two-dimensional linear subspace of the total Hilbert space of the nine physical qubits that is orthogonal to $\mathcal{H}_L$:

$$\sigma_\mu^{(a)} \mathcal{H}_L \perp \mathcal{H}_L, \quad (3.170)$$

for all $a = 1, \dots, 9$ and $\mu = 1, 2, 3$. By measuring the stabilizer operators, an agent thereby projects the quantum information either into the code space $\mathcal{H}_L$ or into one of the error spaces that are orthogonal to $\mathcal{H}_L$. Since this affects both code words $|0\rangle_L$ and $|1\rangle_L$ in the same way, no information about the amplitudes α and β of the logical qubit Eq. (3.168) is obtained. At the same time, based on the information contained in the error syndrome, the logical qubit can easily be rotated back into the code space.

We have not described specific methods for how to measure the operators $M_1, \dots, M_8$, which goes beyond and is not relevant for the current exposition. However, let us close by coming back to our general description of noise and irreversibility at the beginning of this subsection. Equation (3.147) referred to a

situation as depicted in Fig. 3.8, where the quantum information was carried by a single physical qubit. This expression can straightforwardly be generalized to a situation where quantum information is encoded in several physical qubits, e.g., nine qubits in the case of the Shor code, where any of these physical qubits may interact with the environment through a scattering process as in Fig. 3.8. For such a situation, if qubit a is affected, the analogue of Eq. (3.147) has the form

$$|\Psi'\rangle = \sum_{\mu} \left(\sigma_{\mu}^{(a)} |\psi\rangle_L \right) |e_{\mu,a}\rangle, \quad (3.171)$$

with a corresponding set $\{|e_{\mu,a}\rangle\}$ of (unnormalized) environmental states. From the properties of the stabilizer operators and the error spaces as described in this subsection, one can see that the measurement of the observables $M_1, \dots, M_8$ will project the state $|\Psi'\rangle$ into one of the tensor-product terms of the sum, i.e.,

$$|\Psi'\rangle \rightarrow \left(\sigma_{\mu}^{(a)} |\psi\rangle_L \right) |e_{\mu,a}\rangle, \quad (3.172)$$

after which the reduced density matrix of the nine physical qubits, which constitute the memory and carry the logical qubit, is again in a pure state $|\psi'\rangle_L = \sigma_{\mu}^{(a)} |\psi\rangle_L$. That state can easily be transformed back into the original encoded state (Eq. (3.168)) by applying a correction operator $(\sigma_{\mu}^{(a)})^{-1} = \sigma_{\mu}^{(a)}$.¹⁴⁵ If the interaction of the physical qubits with the environment is irreversible, it will “collapse” the state of the logical qubit, according to our discussion in Sect. 3.5.4.1, into one of the terms $\sigma_{\mu}^{(a)} |\psi\rangle_L$. This can also be described by writing the reduced state of the memory as a probabilistic mixture of projectors of the form $\sigma_{\mu}^{(a)} |\psi\rangle_L \langle\psi| \sigma_{\mu}^{(a)}$. The measurement of the stabilizer generators then simply corresponds to finding out in which of the states the memory was left behind by the noise. However one can also think of a situation where the entangling interaction takes place with controlled ancillary degrees of freedom, similar as in the discussion of pre-measurements, delayed choice, and which-way detectors in Sect. 3.3.3. In that case, it is the agent’s measurement that introduces irreversibility and dis-entangles the quantum memory from the ancillary degrees of freedom and thereby generates a random transformation (“error”) on the code space. In both situations, either when the error happens before the syndrome measurement, or when it is only introduced by that very measurement, the trick that preserves the quantum information relies on measuring multi-qubit observables that remove uncertainty about the state of the memory while leaving the superposition of the logical qubit intact.

For a more comprehensive treatment of quantum error correction and stabilizer codes we refer the reader to the literature, e.g., the textbook by Nielsen and Chuang

¹⁴⁵ Technically speaking, the Shor code is a degenerate quantum code, which means that several physical errors, specifically the phase flip errors Z_1, Z_2, Z_3 on the qubits 1,2,3 (similarly for the triplets 4,5,6 and 7,8,9) have the same error syndrome. Even though these errors can thus not be resolved, they can still be corrected because they have the same effect on the code space.

(2010, Ch. 10). In this section we have just given a brief introduction into the basic ideas that describe how an agent can deal with indeterminism, resulting from noisy interactions of a quantum memory with its environment, to protect quantum information. For technological applications, in particular for the realisation of large-scale quantum computers, quantum error correction will be an essential and necessary tool. The development of good quantum error correcting codes is still an active field of research. The specific noise processes relevant for quantum memory will depend on the given hardware implementation as well as on the environment. If one does not precisely know which type of noise may occur under real conditions or if the environment changes over time, one needs flexible and adaptive schemes that go beyond the typical practice of assuming a fixed model of the device and the environment and designing a code on that basis. In Sect. 5.4.1, we will discuss an application of Projective Simulation, as an artificial learning agent, in the context of quantum error correction, using a class of codes known as surface codes. Over and above the simple execution of the error correction protocol for a given surface code, a PS agent can learn how to adapt and change a given code to prevailing conditions in a real implementation.

3.6 Quantum Mechanics and Agency as a Package Deal

The technological harnessing of quantum randomness and entanglement shows how we as agents can make use of QM to expand the range of things we can do in the world. Public device-independent quantum randomness, as envisaged by the NIST beacon project, offers a resource for agents who want certified unpredictability, e.g., to ensure some ideal of fairness; quantum cryptography promises to offer resources for agents who want to be assured that their secret communication cannot be tampered with. In this way, twentieth century philosophical conundra of QM are made useful as resources for sophisticated agents operating in our twenty-first century technological world.

The match between agency and quantum mechanics, however, runs even deeper, in two specific ways. One point, which we will describe in Sect. 3.6.2, is that to accept QM as an indeterministic theory means to embrace a world-view on which ontological emergence of entities such as agents is really possible. It is not, of course, that QM guarantees that agency will actually emerge: We have good reasons to assume that QM is true of distant solar systems in which there is neither life nor agency, and perhaps but for some coincidences in the history of the Earth, our solar system would have been one of them. But base-level indeterminism, as offered by QM, is a necessary condition for the emergence of agency. In this way, QM supports agency. The other point, which we will describe in the following Sect. 3.6.1, is that there is a crucial lesson that Bell's theorem teaches about the alignment of QM and agency that has to do with the very idea of a scientific experiment.

3.6.1 Lessons from Bell's Theorem

Bell's theorem, together with the empirical violation of Bell's inequalities, provides very good evidence that a future theory that can account for the quantum phenomena that we already know about will *either* adhere to a basic probabilistic structure, *or* it will have to do away with our notions of both agency and experiment. We are facing a choice between two world-views that implies a package deal: On the one hand, *if there are agents doing experiments, then randomness in nature is fundamental*. On the other hand, *if we reject the fundamental indeterminism of QM, then we have to embrace agency scepticism*.

This bifurcation comes about as follows. On the one hand, there is always the option that above we called *agency scepticism* (see Sect. 2.4.2): to take on a non-perspective “from nowhere” according to which things simply happen, with or without explanation (cf. Nagel, 1986, 114). Some of those happenings are what *we* (from now-here, rather than from nowhere) would describe as agents making choices—for example, choices about which experiment to perform on a quantum system and thus, which observable to consider regarding that system in its given state. But “viewed from nowhere” these choices are just happenings in nature like any others, and they are not special in any way, because nothing is. So a full theory of what happens in the world as a whole should also apply to these happenings and thereby incorporate them as a matter of natural fact. Whether these happenings can be integrated into a comprehensive *deterministic* picture is then more a matter of pragmatics than one of principle. The real course of history is always a unique, linear one, one thing following another; it is conceptually impossible that two incompatible possibilities should be realised together.¹⁴⁶ So from the structure of what happens, a deterministic story is always logically possible; one just has to claim that the actual course of events was also the only possibility that was ever there, so that the future was never open. No observation can decisively rule out this assessment on empirical grounds—even if extraordinary, unexpected things happen, one can always retreat to the (self-insulating) idea that in the whole pre-ordained order of things, this was simply what was going to happen (“Que sera, sera”,¹⁴⁷ as the mock-Spanish proverb goes). The open question from the point of view of scientific theory building then is only whether a compact, rule-based description of the actual sequence of what happens can be found—whether the sequence of states at different times evolves according to some clear and simple

¹⁴⁶ This assessment stays in place if we replace the pre-relativistic notion of successive states of the whole universe by a relativistic space-time manifold. Whatever happens, exactly one such manifold is successively realised. In fact, the idea of Minkowski space-time as a container for all that happens has itself given rise to several arguments in favour of determinism. See Rietdijk (1966) and Putnam (1967) for two early and influential arguments that draw such conclusions. These arguments are, however, not conclusive. See Belnap et al. (2022) for a formally worked out theory of indeterminism in relativistic space-time.

¹⁴⁷ The title of a famous Doris Day song; see Hartman (2013) for the faux-Roman etymology.

deterministic rule. That may or may not be so, depending to a significant extent on which types of rules are admitted. Finding such rules is actually a research program, under the heading of *superdeterminism*. The “super” in “superdeterminism” pays tribute to the fact that in such a theory, there are no free choices in experiments any more. Doing an experiment is a pattern in the overall evolution of the universe like any other; it is fully integrated into the global picture. On superdeterminism, no agent ever intervenes in the course of nature, because that course of events already prescribes what any agent does. As we already indicated, we do not find this to be a scientifically warranted approach, and it is certainly not forced upon us for any empirical reasons—standard QM works just fine. Reasons for the view can only be brought forth by agents who want to defend their metaphysical preference for determinism—thereby curiously claiming leeway for their own theory formation within a world-view that excludes such leeway for all agents, and which even implies that there cannot be agents at all, as a matter of principle. At any rate, Bell’s theorem does not significantly constrain such theories, because the independent and free selection of measurement settings, a crucial premise of Bell’s theorem, is rejected.¹⁴⁸

The other path at the mentioned bifurcation—the one we believe gets things right—is to affirm that there are agents in the world who build models and theories as tools to suit their needs and who can make choices—for example, which experiments to perform in the lab. One has to accept that on this view, even a possible successor to QM can make probabilistic predictions only, because there is fundamental indeterminism in nature: the future is truly open. As we argued in Chap. 2, such base-level indeterminism is required for the existence of real ontological levels in nature and thus, for the real existence of agents in the first place. That is a nice indication of the self-consistency of the view. The price to pay for a world in which agents can make a real difference to what happens is that the world has to leave enough slack in terms of open possibilities for these agents to make such differences. This does not seem to be asking too much.

3.6.2 *QM as a Basis for the Ontological Emergence of Agency*

Basic indeterminism of course does not guarantee agential freedom. But it provides a material basis with the characteristics that are necessary for the real ontological

¹⁴⁸ See Müller and Placek (2023) for some pertinent comments and references to the literature. To mention one prominent example of a defender of superdeterminism, Nobel laureate t’Hooft (2016) explicitly affirms that view and a bitmap picture of all happenings in nature in his mathematically detailed “cellular automaton interpretation” of QM. By negating item (iv) of our explication of agency given in Sect. 2.6.1, he also subscribes to agency scepticism. On his view, there are only observers, and their would-be actions are indeed the inevitable consequences of what has gone before: “the observers themselves are controlled by deterministic laws. They certainly have no ‘free will’, their actions all have roots in the past” (t’Hooft, 2016, 4).

emergence of complex entities, including agents. Quantum holism functions as a prelude for emergence.

From Quantum Holism to the Emergence of Novel Ontological Levels As we mentioned in the previous Chap. 2, entangled quantum systems are simple, yet clear instances of the phenomenon of holism, according to the definition that a system consisting of parts is holistic if and only if its state cannot be derived as a function of the state of the parts. For the Bell singlet state (Eq. (3.102)), the state of the parts is the completely mixed state $\rho = \frac{1}{2}\hat{1}$, both for Alice and for Bob, and these states of the parts are compatible with many joint states, ranging from the completely mixed joint state to the Bell states of Eq. (3.125). Thus, the state of the whole system is not a function of the states of the parts. Holism in this sense is, as we said, pervasive in QM. Insofar as holism is sufficient for epistemic emergence, then, QM offers a plethora of instances of epistemic emergence.

As we argued in Sect. 2.2.3.4 above, however, agency instantiates a stronger principle: agents must be *ontologically* emergent entities. Ontological emergence is required for the existence of all higher ontological levels in nature, which on Earth include organised life, agency, and human agency. As we said, we are realists about entities at these higher levels, and we have already indicated that we believe that QM is compatible with a world view that encompasses such entities. In this final part of our Chapter on QM, we want to sketch an argument for this compatibility claim, mostly by way of addressing a crucial criticism against the view: How can agential control reach down through various ontological levels to their quantum mechanical base?

Moving One's Body When describing ontological emergence in Sect. 2.2.3.4, we said that an ontologically emergent entity organises its matter as material, and that this organisation implies a kind of top-down control. In human agency, this is easy to make sense of. As an agent, I can move my body. For example, I can move my hand to drink from the cup in front of me, thereby settling the open future with respect to the question of where my hand will be in the next few seconds. The future is open in that respect because the physiological organisation of my body leaves open a wide range of possibilities for the future arrangement of my limbs: My limbs are connected by joints, and these allow for many different positions. It is up to me as an agent to settle matters concerning the position of my limbs within the range left open by the circumstances of the case. When I am tied up, that range is very small; when I am unconscious, my agency is suspended and that range is zero. But in normal circumstances, that range is such that it includes both my gripping a cup in front of me to drink from it *and* refraining from doing so, e.g., by keeping still. I can normally move my limbs. In this sense, as an agent I have top-down control over their movement.

The lee-way in the future position of my limbs, as organs of my body, is made possible in turn by *their* internal organisation: My muscles, nerves, sinews, cartilages and the other relevant parts of my body are connected in such a way as to allow for a wide range of possible configurations, and their configuration at one time leaves open many different possible configurations at later times. In moving my

body, I am also making use of that deeper-level openness. The movement of my hand settles many of the possibilities for its parts, which include lots of small muscles. It does not seem right to say that *I as an agent* settle these matters concerning my physiology directly, at least not as far as most of the details are concerned. It is true that I am in agential control of some of my skeletal muscles, directly or indirectly: I can flex or relax some of them voluntarily, and I can use mental imagery to influence the state of some others, e.g., in muscle relaxation exercises.¹⁴⁹ But certainly the state of my individual muscle cells is beyond my reach as an agent (unless I affix electrodes to them to stimulate them). The same holds true for my nerve cells and the neurons in my motor cortex that influence them. Thus in normal circumstances, it is really the motion of my limbs (which I as an agent *do* control, by moving them) that settles matters at that lower level of the individual muscle cells and neurons.

Control All the Way Down? So far, so good, but a bit further down we may seem to get into trouble. The settling of the state of my muscle cells comes by way of the settling of the state of the nerve cells that innervate the respective muscle, and at the cellular level, this translates into the polarisation and depolarisation of the membranes of the individual neurons that make up the nerve, and of the relevant neurons in the motor cortex. That is, *their* polarisation has to be settled as part of the process that is *my* (agential) moving of my hand. Now we know that the polarisation and depolarisation of cell membranes comes down to the collective opening and closing of large numbers of ion channels in these membranes, and ions consequently diffusing or being blocked from diffusing through them. Each such ion channel is a large molecule that can assume different three-dimensional configurations that allow for the passage of ions through the membrane or block that passage. And the transformation of those molecules' three-dimensional structure from one configuration to another is, to a good approximation, a measurement-like irreversible process at the quantum level. But this in turn means that the Born rule applies to that transformation, and that it is, therefore, fundamentally random. How could it be controlled?

There seems to be a possibly fatal tension here: Does the agent who is moving her hand have to influence, albeit indirectly, the outcome of random processes at the quantum level? That would be bad news, because such an influence is impossible if the relevant quantum dynamics is indeed fundamentally random, and we just spent a significant number of pages arguing that is so. Do we have to take recourse to a further unanalysable notion of agent causation here,¹⁵⁰ or conclude that all these assumptions about top-down influences are simply incoherent?

¹⁴⁹ It is interesting to observe that the relevant mental imagery may include imagined motions that I know to be absolutely impossible, but which still have an effect. Thus, for example, in a yoga class I may be instructed to breathe through my navel, which everybody in the room knows to be impossible, but trying to do exactly that still works as a means of controlling certain muscles in my body that are otherwise hard to influence.

¹⁵⁰ See Clarke et al. (2021, §3) for an overview of the notion. See Müller (2022) and Morales (2023) for two recent contributions to the debate about agent causation.

Adaptive Control at the Quantum Level The quantum level is a level at which pure randomness plays out. If Alice has prepared an atom in the $|\uparrow, z\rangle$ spin state and feeds it into a Stern-Gerlach device oriented along the x -axis, there is nothing that she or anybody else can do to influence which channel the atom comes out—the spin alignment is a fundamentally indeterministic 50:50 chance process. The worry about agential control is that for this reason, our idea of level building by matter being organised as material cannot get off the ground, because nothing can organise the quantum events at the bottom of the chain of levels.

But this would be too quick. The quantum level is not one of chaos. Quantum indeterminism is fundamental, but it is tuneable. We have already seen that directed influences are possible and in fact technologically useful. As we showed in discussing the dynamics of spin alignment above, Alice *can* do something. For example, she can switch on a B -field, as discussed in Sect. 3.3.3, and tune it such that her $|\uparrow, z\rangle$ atom will be guaranteed to come out of the upper channel of her x -aligned device. This is what Eq. (3.83) tells us; that equation also prescribes quantitatively the magnetic field that Alice needs to choose. Thus, Alice can get the z -prepared atom to come out of the upper channel of her x -aligned apparatus after all. But she achieves this not by a “brute force” intervention into the chance process of spin measurement, but by adaptively controlling the atom’s environment. The modulation of the environment is a way of top-down control reaching into the quantum level without breaking any laws, and without suspending the Born rule.

So far, this is just an indication of how control and quantum randomness can go together. But we will see in Chap. 5 how the agency model of Projective Simulation allows us to fill in missing details here. Agency can be based on a lowest level like QM at which random processes are fundamental, given that there are ways of tuning the environment in which such processes take place. In fact, the basic learning rule of classical PS describes such changes—updates of the weights for a biased random walk—in a classical environment, and PS can be realised in a quantum environment as well. In Sect. 5.6 we will show via a concrete model how the worry about top-down control in a quantum environment can be dismissed: a quantum version of PS shows how the tuning of the environment in which random quantum processes unfold can be used as a model for learning and agency.

References

- Abellán, C., Acín, A., Alarcón, A., & The BIG Bell Test Collaboration. (2018). Challenging local realism with human choices. *Nature*, 557, 212–216.
- Adelberger, E. G., García, A., Robertson, R. G. H., Snover, K. A., Balantekin, A. B., Heeger, K., Ramsey-Musolf, M. J., Bemmerer, D., Junghans, A., Bertulani, C. A., Chen, J.-W., Costantini, H., Prati, P., Couder, M., Überseder, E., Wiescher, M., Cyburt, R., Davids, B., Freedman, S. J., Gai, M., Gazit, D., Gialanella, L., Imbriani, G., Greife, U., Hass, M., Haxton, W. C., Itahashi, T., Kubodera, K., Langanke, K., Leitner, D., Leitner, M., Vetter, P., Winslow, L., Marcucci, L. E., Motobayashi, T., Mukhamedzhanov, A., Tribble, R. E., Nollett, K. M., Nunes, F. M., Park, T.-S., Parker, P. D., Schiavilla, R., Simpson, E. C., Spitaleri, C., Strieder, F., Trautvetter,

- H.-P., Suemmerer, K., & Typel, S. (2011). Solar fusion cross sections. II. The pp chain and CNO cycles. *Reviews of Modern Physics*, 83, 195–245.
- Agresti, I., Poderini, D., Guerini, L., Mancusi, M., Carvachol, G., Aolita, L., Cavalcanti, D., Chaves, R., & Sciarrino, F. (2020). Experimental device-independent certified randomness generation with an instrumental causal structure. *Communications Physics*, 3, 10.
- Alicki, R., & Lendi, K. (2007). *Quantum dynamical semigroups and applications*. Lecture Notes in Physics (Vol. 717). Springer.
- Anscombe, G. E. M. (1971). *Causality and determination*. Cambridge University Press. Page references are to the reprint in her *Collected Philosophical Papers*, Vol. 2, Oxford 1981, pp 133–147.
- Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, 49, 1804–1807.
- Atkinson, D., & Peijnenburg, J. (2022). How certain is Heisenberg's uncertainty principle? *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, 12(1), 1–21.
- Auletta, G., Fortunato, M., & Parisi, G. (2009). *Quantum mechanics*. Cambridge University Press.
- Bagan, E., Bergou, J. A., Cottrell, S. S., & Hillery, M. (2016). Relations between coherence and path information. *Physical Review Letters*, 116, 160406.
- Bahcall, J. N., & May, R. M. (1969). The rate of the proton-proton reaction and some related reactions. *The Astrophysical Journal*, 155, 501.
- Ballentine, L. E. (2015). *Quantum mechanics. A modern development* (2nd ed.). World Scientific.
- Barnett, S. M. (2009). *Quantum information*. Oxford Master Series in Atomic, Optical, and Laser Physics. Oxford University Press.
- Bauer, P., Stevens, B., & Hazeleger, W. (2021). A digital twin of Earth for the green transition. *Nature Climate Change*, 11, 80–83.
- Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, 1, 195–200. Reprinted as Ch. 2 of Bell's *Speakable and Unspeakeable in Quantum Mechanics* (1987).
- Bell, J. S. (1966). On the problem of hidden variables in quantum mechanics. *Reviews of Modern Physics*, 38, 447–452. Reprinted as Ch. 1 of Bell's *Speakable and Unspeakeable in Quantum Mechanics* (1987).
- Bell, J. S. (1987). *Speakable and unspeakeable in quantum mechanics*. Cambridge University Press.
- Bell, J. S. (1990). Against 'measurement'. *Physics World*, 3, 33–40.
- Belnap, N., Müller, T., & Placek, T. (2022). *Branching space-times. theory and applications*. Oxford studies in philosophy of science. Oxford University Press.
- Bengtsson, I., & Życzkowski, K. (2017). *Geometry of quantum states: An introduction to quantum entanglement* (2nd ed.). Cambridge University Press.
- Bennett, C. H., & Wiesner, S. J. (1992). Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Physical Review Letters*, 69, 2881–2884.
- Bennett, C. H., Brassard, G., Crépeau, C., Jozsa, R., Peres, A., & Wootters, W. K. (1993). Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters*, 70, 1895–1899.
- Berryman, S. (2022). Ancient atomism. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2022 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2022/entries/atomism-ancient/>.
- Bierhorst, P., Knill, E., Glancy, S., Zhang, Y., Mink, A., Jordan, S., Rommal, A., Liu, Y.-K., Christensen, B., Nam, S. W., et al. (2018). Experimentally generated randomness certified by the impossibility of superluminal signals. *Nature*, 556(7700), 223–226.
- Black, M. (1946). *Critical thinking. An introduction to logic and scientific method*. Prentice Hall.
- Bluvstein, D., Levine, H., Semeghini, G., Wang, T. T., Ebadi, S., Kalinowski, M., Keesling, A., Maskara, N., Pichler, H., Greiner, M., Vuletić, V., & Lukin, M. D. (2022). A quantum processor based on coherent transport of entangled atom arrays. *Nature*, 604(7906), 451–456.
- Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of "hidden" variables. I. *Physical Review*, 85, 166–179.
- Bohr, N. (1934). *Atomic theory and the description of nature*. Cambridge University Press.

- Briegel, H. J. (2009). Cluster states. In D. Greenberger, K. Hentschel, & F. Weinert (Eds.), *Compendium of quantum physics* (pp. 96–105). Springer.
- Briegel, H. J., & Raussendorf, R. (2001). Persistent entanglement in arrays of interacting particles. *Physical Review Letters*, 86, 910–913.
- Briegel, H.-J., Dür, W., Cirac, J. I., & Zoller, P. (1998). Quantum repeaters: The role of imperfect local operations in quantum communication. *Physical Review Letters*, 81, 5932–5935.
- Busch, P., Lahti, P. J., & Mittelstaedt, P. (1996). *The quantum theory of measurement* (Vol. 2, 2nd revised ed.). Lecture notes in physics monographs. Springer.
- Butterfield, J. (2005). Determinism and indeterminism. In *Routledge encyclopedia of philosophy* (Vol. 3). Routledge.
- Cabello, A., Estebaranz, J., & García-Alcaine, G. (1996). Bell-Kochen-Specker theorem: A proof with 18 vectors. *Physics Letters A*, 212(4), 183–187.
- Calsamiglia, J., Hartmann, L., Dür, W., & Briegel, H.-J. (2005). Spin gases: Quantum entanglement driven by classical kinematics. *Physical Review Letters*, 95, 180502.
- Carlson, J., Riska, D. O., Schiavilla, R., & Wiringa, R. B. (1991). Weak proton capture reactions on ^1H and ^3He and tritium β decay. *Physical Review C*, 44(2), 619–625.
- Chiribella, G., D’Ariano, G. M., & Perinotti, P. (2009). Theoretical framework for quantum networks. *Physical Review A*, 80, 022339.
- Christensen, N., & Meyer, R. (2022). Parameter estimation with gravitational waves. *Reviews of Modern Physics*, 94, 025001.
- Clarke, R., Capes, J., & Swenson, P. (2021). Incompatibilist (nondeterministic) theories of free will. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2021/entries/incompatibilism-theories/>
- Cohen-Tannoudji, C., Diu, B., & Laloë, F. (2019). *Quantum mechanics: Volume I: Basic concepts, tools, and applications* (2nd ed.). Wiley-VCH.
- Commins, E. D. (2012). Electron spin and its history. *Annual Review of Nuclear and Particle Science*, 62(1), 133–157.
- Cronin, A. D., Schmiedmayer, J., & Pritchard, D. E. (2009). Optics and interferometry with atoms and molecules. *Reviews of Modern Physics*, 81, 1051–1129.
- Dennett, D. C. (2003). *Freedom Evolves*. Vinking Books.
- Diaconis, P., Holmes, S., & Montgomery, R. (2007). Dynamical bias in the coin toss. *SIAM Review*, 49(2), 211–235.
- Dieks, D. (1982). Communication by EPR devices. *Physics Letters A*, 92(6), 271–272.
- Diestel, R. (2017). *Graph theory*. Graduate texts in mathematics (Vol. 173). Springer.
- Ducheyne, S. (2014). Newton on action at a distance. *Journal of the History of Philosophy*, 52(4), 675–701.
- Earman, J. (1986). *A primer on determinism*. D. Reidel.
- Einstein, A., & Ehrenfest, P. (1922). Quantentheoretische Bemerkungen zum Experiment von Stern und Gerlach. *Zeitschrift für Physik*, 11, 31–34.
- Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47(10), 777–780.
- Ekert, A. (1991). Quantum cryptography based on Bell’s theorem. *Physical Review Letters*, 67, 661–663.
- Englert, B.-G. (1996). Fringe visibility and which-way information: An inequality. *Physical Review Letters*, 77, 2154.
- Englert, B.-G. (2006a). *Lectures on quantum mechanics. Volume 1: Basic matters*. World Scientific.
- Englert, B.-G. (2006b). *Lectures on quantum mechanics. Volume 3: Perturbed evolution*. World Scientific.
- Englert, B.-G. (2013). On quantum theory. *European Physical Journal D*, 67, 238.
- Englert, B.-G., Schwinger, J., & Scully, M. O. (1988). Is spin coherence like Humpty-Dumpty? I. Simplified treatment. *Foundations of Physics*, 18, 1045–1056.
- Feynman, R. P., Leighton, R. B., & Sands, M. I. (1989). *The feynman lectures on physics* (Vol. 3, Commemorative ed.). Addison-Wesley.

- Forman, P. (1968). The doublet riddle and atomic physics circa 1924. *Isis*, 59, 156–174.
- Frauchiger, D., & Renner, R. (2018). Quantum theory cannot consistently describe the use of itself. *Nature Communications*, 9, 3711.
- French, S. (2019). Identity and individuality in quantum theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2019/entries/qt-idind/>.
- Friedrich, B. (2023). A century ago the Stern-Gerlach experiment ruled unequivocally in favor of quantum mechanics. *Israel Journal of Chemistry*, 63, e202300047.
- Friedrich, B., & Herschbach, D. (2003). Stern and Gerlach: How a bad cigar helped reorient atomic physics. *Physics Today*, 56, 53–59.
- Frisch, R., & Segré, E. (1933). Über die Einstellung der Richtungsquantelung. II. *Zeitschrift für Physik*, 80, 610–616.
- Gachechiladze, M., Budroni, C., & Gühne, O. (2016). Extreme violation of local realism in quantum hypergraph states. *Physical Review Letters*, 116, 070401.
- Galindo, A., & Pascual, P. (1990). *Quantum mechanics I*. Theoretical and mathematical physics. Springer.
- Gerlach, W., & Stern, O. (1922). Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld. *Zeitschrift für Physik*, 9, 349–352.
- Giustina, M., Versteegh, M. A. M., Wengerowsky, S., Handsteiner, J., Hochrainer, A., Phelan, K., Steinlechner, F., Kofler, J., Larsson, J.-A., Abellán, C., Amaya, W., Pruneri, V., Mitchell, M. W., Beyer, J., Gerrits, T., Lita, A. E., Shalm, L. K., Nam, S. W., Scheidl, T., Ursin, R., Wittmann, B., & Zeilinger, A. (2015). Significant-loophole-free test of Bell's theorem with entangled photons. *Physical Review Letters*, 115, 250401.
- Gleason, A. M. (1957). Measures on the closed subspaces of a Hilbert space. *Indiana University Mathematics Journal*, 6(4), 885–893.
- Goldstein, H., Poole, C. P., & Safko, J. L. (2002). *Classical mechanics* (3rd ed.). Addison-Wesley.
- Gottesman, D. (1997). Stabilizer codes and quantum error correction. arXiv quant-ph/9705052.
- Goudsmit, S. A. (1971). De ontdekking van de electronenrotatie. *Nederlands Tijdschrift voor Natuurkunde*, 37(16), 386–392. English translation as “The discovery of the electron spin” (by J.H. van der Waals) in *Foundations of Modern EPR*, ed. G.R. Eaton, S.S. Eaton, and K.M. Salikhov, World Scientific, Singapore, 1998, pp. 1–12.
- Grassl, M., Klappenecker, A., & Rötteler, M. (2002). Graphs, quadratic forms, and quantum codes. *Proceedings of the international symposium on information theory, Lausanne, June/July 2002*. arXiv quant-ph/0703112.
- Greenberger, D. M., & Yasin, A. (1988). Simultaneous wave and particle knowledge in a neutron interferometer. *Physics Letters A*, 128(8), 391–394.
- Greenberger, D. M., Horne, M. A., Shimony, A., & Zeilinger, A. (1990). Bell's theorem without inequalities. *American Journal of Physics*, 58, 1131–1143.
- Gröblacher, S., Paterek, T., Kaltenbaek, R., Brukner, Č., Żukowski, M., Aspelmeyer, M., & Zeilinger, A. (2007). An experimental test of non-local realism. *Nature*, 446(7138), 871–875.
- Gühne, O., Tóth, G., Hyllus, P., & Briegel, H. J. (2005). Bell inequalities for graph states. *Physical Review Letters*, 95, 120405.
- Gustafson, S. J., & Sigal, I. M. (2011). *Mathematical concepts of quantum mechanics*. Universitext. Springer.
- Hacking, I. (1983). *Representing and intervening*. Cambridge University Press.
- Hacking, I. (1990). *The taming of chance*. Cambridge University Press.
- Häffner, H., Roos, C., & Blatt, R. (2008). Quantum computing with trapped ions. *Physics Reports*, 469, 155–203.
- Hartman, L. (2013). “Que sera sera”: The English roots of a pseudo-Spanish proverb. *Proverbium*, 30, 51–104.
- Hayes, W. B. (2007). Is the outer solar system chaotic? *Nature Physics*, 3, 689–691.
- Healey, R. (2023). Quantum-Bayesian and pragmatist views of quantum theory. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (winter 2023 ed.).

- Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2023/entries/quantum-bayesian/>.
- Hein, M., Eisert, J., & Briegel, H. J. (2004). Multiparty entanglement in graph states. *Physical Review A*, 69, 062311.
- Hein, M., Dür, W., Eisert, J., Raussendorf, R., Van den Nest, M., & Briegel, H. J. (2006). Entanglement in graph states and its applications. In G. Casati, D. Shepelyansky, P. Zoller, & G. Benenti (Eds.), *Quantum computers, algorithms and chaos*. Proceedings of the international school of physics “Enrico Fermi” (Vol. 162, pp. 115–218). IOS Press. arXiv quant-ph/0602096.
- Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43, 172–198.
- Hensen, B., Bernien, H., Dréau, A., Reiserer, A., Kalb, N., Blok, M., Ruitenberg, J., Vermeulen, R., Schouten, R., Abellán, C., et al. (2015). Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526(7575), 682–686.
- Hentschel, K. (1996). Die Entdeckung des Zeeman-Effekts. *Physikalische Blätter*, 52(12), 1232–1235.
- Herbert, N. (1982). FLASH—A superluminal communicator based upon a new kind of quantum measurement. *Foundations of Physics*, 12(12), 1171–1179.
- Herrero-Collantes, M., & Garcia-Escartin, J. C. (2017). Quantum random number generators. *Reviews of Modern Physics*, 89, 015004.
- Hilgevoord, J., & Uffink, J. (2024). The Uncertainty Principle. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2024 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2024/entries/qt-uncertainty/>.
- Hornberger, K., Gerlich, S., Haslinger, P., Nimmrichter, S., & Arndt, M. (2012). Colloquium: Quantum interference of clusters and molecules. *Reviews of Modern Physics*, 84, 157–173.
- Horodecki, R., Horodecki, P., Horodecki, M., & Horodecki, K. (2009). Quantum entanglement. *Reviews of Modern Physics*, 81, 865–942.
- Hurwitz, R. (1992). Milestones in magnetic resonance: Rabi et al. (1938), ‘A new method of measuring nuclear magnetic moment’. *Journal of Magnetic Resonance Imaging*, 2(2), 131–133.
- IDQ (2022). ID Quantique website. <https://www.idquantique.com/>.
- Ismael, J. (2025). Quantum mechanics. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2025 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2025/entries/qm/>.
- Jaksch, D., Briegel, H.-J., Cirac, J. I., Gardiner, C. W., & Zoller, P. (1999). Entanglement of atoms via cold controlled collisions. *Physical Review Letters*, 82, 1975–1978.
- Kochen, S., & Specker, E. (1967). The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics*, 17, 59–87.
- Kuhn, T. S. (1962). *The structure of scientific revolutions* (4th ed.). University of Chicago Press. With an introduction by Ian Hacking, 2012.
- Laplace, P. S. (1820). *Théorie Analytique des Probabilités*. V. Courcier. English translation by F. W. Truscott and F. L. Emory: *A philosophical essay on probabilities*, New York, NY: Dover, 1951. Page references are to this translation.
- Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48(1), 19–49.
- Lewis, D. K. (1986a). *On the plurality of worlds*. Blackwell.
- Lewis, D. K. (1986b). *Philosophical Papers* (Vol. 2). Oxford University Press.
- Libbrecht, K. G. (2021). *Snow crystals: A case study in spontaneous structure formation*. Princeton University Press.
- Lidar, D. A., & Brun, T. A. (Eds.). (2013). *Quantum error correction*. Cambridge University Press.
- Lindblad, G. (1976). On the generators of quantum dynamical semigroups. *Communications in Mathematical Physics*, 48(2), 119–130.
- Linden, N., & Popescu, S. (1998). On multi-particle entanglement. *Fortschritte der Physik*, 46(4–5), 567–578.
- Lockwood, M. (1996). ‘Many minds’ interpretations of quantum mechanics. *The British Journal for the Philosophy of Science*, 47(2), 159–188.

- Ma, X.-S., Kofler, J., & Zeilinger, A. (2016). Delayed-choice gedanken experiments and their realizations. *Reviews of Modern Physics*, 88, 015005.
- Mackenzie, D. (2009). The fifty-one percent solution. In D. Mackenzie (Ed.), *What's happening in the mathematical sciences?* (Vol. 7, pp. 34–45). American Mathematical Society.
- Majumder, A., Krumm, M., Radkohl, T., Poulsen Nautrup, H., Jerbi, S., & Briegel, H. J. (2024). Variational measurement-based quantum computation for generative modeling. *Physical Review A*, 110, 062616.
- Mannalath, V., Mishra, S., & Pathak, A. (2023). A comprehensive review of quantum random number generators: Concepts, classification and the origin of randomness. *Quantum Information Processing*, 22, 439.
- Margalit, Y., Dobkowski, O., Zhou, Z., Amit, O., Japha, Y., Moukouri, S., Rohrich, D., Mazumdar, A., Bose, S., Henkel, C., & Folman, R. (2021). Realization of a complete Stern-Gerlach interferometer: Toward a test of quantum gravity. *Science Advances*, 7(22), eabg2879.
- Mermin, N. D. (1981). Quantum mysteries for anyone. *Journal of Philosophy*, 78(7), 397–408.
- Mermin, N. D. (1990). Quantum mysteries revisited. *American Journal of Physics*, 58, 731–734.
- Michelson, A. A. (1903). *Light waves and their uses*. University of Chicago Press.
- Miller, J., Sanders, S., & Miyake, A. (2017). Quantum supremacy in constant-time measurement-based computation: A unified architecture for sampling and verification. *Physical Review A*, 96, 062320.
- Morales, J. (2023). Emergent agent causation. *Synthese*, 201, 138.
- Müller, T. (2022). “Let’s build an Anscombe box”: assessing Anscombe’s rebuttal of the statistics objection against indeterminism-based free agency. *Synthese*, 200, 74.
- Müller, T., & Placek, T. (2023). On the condition of setting independence. *European Journal for the Philosophy of Science*, 13, 53.
- Myrvold, W. (2022). Philosophical Issues in Quantum Theory. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2022/entries/qt-issues/>.
- Myrvold, W., Genovese, M., & Shimony, A. (2024). Bell’s Theorem. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2024 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2024/entries/bell-theorem/>.
- Nagel, T. (1986). *The view from nowhere*. Oxford University Press.
- Nielsen, M. A., & Chuang, I. L. (2010). *Quantum computation and quantum information* (10th anniversary ed.). Cambridge University Press.
- NIST (2018). NIST’s new quantum method generates really random numbers. <https://www.nist.gov/news-events/news/2018/04/nists-new-quantum-method-generates-really-random-numbers>.
- NIST (2022). <https://csrc.nist.gov/projects/interoperable-randomness-beacons>.
- Norton, J. D. (2008). The Dome: An unexpectedly simple failure of determinism. *Philosophy of Science*, 75(5), 786–798.
- Obenchain, T. G. (2016). *Genius belabored: Childbed fever and the tragic life of ignaz semmelweis*. The University of Alabama Press.
- Paraniak, M. M., & Englert, B.-G. (2021). Quantum dynamical simulation of a transversal Stern-Gerlach interferometer. *Symmetry*, 13(9), 1660.
- Pauli, W. (1925). Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Physik*, 31, 765–783.
- Pauli, W. (1927). Zur Quantenmechanik des magnetischen Elektrons. *Zeitschrift für Physik*, 43, 601–623.
- Pavičić, M., Merlet, J.-P., McKay, B., & Megill, N. D. (2005). Kochen-Specker vectors. *Journal of Physics A: Mathematical and General*, 38(7), 1577.
- Pechukas, P. (1994). Reduced dynamics need not be completely positive. *Physical Review Letters*, 73, 1060–1062.
- Peres, A. (1978). Unperformed experiments have no results. *American Journal of Physics*, 46, 745–747.

- Peres, A. (1991). Two simple proofs of the Kochen-Specker theorem. *Journal of Physics A: Mathematical and General*, 24(4), L175.
- Peres, A. (2002). *Quantum theory: Concepts and methods*. Fundamental theories of physics (Vol. 57). Kluwer Academic Publishers.
- Phipps, T. E., & Stern, O. (1932). Über die Einstellung der Richtungsquantelung. *Zeitschrift für Physik*, 73, 185–191.
- Pironio, S., Acín, A., Massar, S., de la Giroday, A. B., Matsukevich, D. N. and 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.
- Putnam, H. (1967). Time and physical geometry. *The Journal of Philosophy*, 64(8), 240–247.
- Rabi, I. I., Zacharias, J. R., Millman, S., & Kusch, P. (1938). A new method of measuring nuclear magnetic moment. *Physical Review*, 53, 318.
- Rabi, I. I., Millman, S., Kusch, P., & Zacharias, J. R. (1939). The molecular beam resonance method for measuring nuclear magnetic moments. The magnetic moments of ${}^3\text{Li}^6$, ${}^3\text{Li}^7$ and ${}^9\text{F}^{19}$. *Physical Review*, 55, 526–535.
- Ragab, H., Milburn, A., Razavi, K., Bos, H., & Giuffrida, C. (2021). Crosstalk: Speculative data leaks across cores are real. In *2021 IEEE symposium on security and privacy (SP)* (pp. 1852–1867).
- Rauch, H. (1998). Quantum phenomena studied by neutron interferometry. *Physica Scripta*, 1998(T76), 24.
- Rauch, H., Zeilinger, A., Badurek, G., Wilfing, A., Bauspiess, W., & Bonse, U. (1975). Verification of coherent spinor rotation of fermions. *Physics Letters A*, 54(6), 425–427.
- Raussendorf, R., & Briegel, H. J. (2001). A one-way quantum computer. *Physical Review Letters*, 86, 5188–5191.
- Raussendorf, R., Browne, D. E., & Briegel, H. J. (2003). Measurement-based quantum computation on cluster states. *Physical Review A*, 68, 022312.
- Rietdijk, C. (1966). A rigorous proof of determinism derived from the special theory of relativity. *Philosophy of Science*, 33(4), 341–344.
- Rumberg, A. (2016). *Transitions toward a semantics for real possibility*. Ph.D. Thesis, University of Utrecht.
- Russell, S., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education.
- Saffman, M. (2016). Quantum computing with atomic qubits and Rydberg interactions: progress and challenges. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(20), 202001.
- Saggio, V., Asenbeck, B. E., Hamann, A., Strömberg, T., Schiansky, P., Dunjko, V., Friis, N., Harris, N. C., Hochberg, M., Englund, D., Wölk, S., Briegel, H. J., & Walther, P. (2021). Experimental quantum speed-up in reinforcement learning agents. *Nature*, 591(7849), 229–233.
- Sakurai, J. J., & Napolitano, J. (2020). *Modern quantum mechanics* (3rd ed.). Cambridge University Press.
- Schindler, P., Nigg, D., Monz, T., Barreiro, J. T., Martinez, E., Wang, S. X., Quint, S., Brandl, M. F., Nebendahl, V., Roos, C. F., Chwalla, M., Hennrich, M., & Blatt, R. (2013). A quantum information processor with trapped ions. *New Journal of Physics*, 15(12), 123012.
- Schlingemann, D. (2002). Stabilizer codes can be realized as graph codes. *Quantum Information and Computation*, 2(4), 307–323.
- Schmidt-Böcking, H., Schmidt, L., Lüdde, H. J., Trageser, W., Templeton, A., & Sauer, T. (2016). The Stern-Gerlach experiment revisited. *The European Physical Journal H*, 41, 327–364.
- Schmidt-Böcking, H., Gruber, G., & Friedrich, B. (2023). One hundred years ago Alfred Landé unriddled the anomalous Zeeman effect and presaged electron spin. *Physica Scripta*, 98(1), 014005.
- Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, 23, 807–812, 823–828, 844–849. English translation by J. D. Trimmer, “The Present Situation in Quantum Mechanics”, *Proceedings of the American Philosophical Society* 124(5), 1980, pp. 323–338.

- Schwinger, J. (2001). In B.-G. Englert (Ed.), *Quantum mechanics. Symbolism of atomic measurements*. Springer.
- Schwinger, J., Scully, M. O., & Englert, B.-G. (1988). Is spin coherence like Humpty-Dumpty? II. General theory. *Zeitschrift für Physik D. Atoms, Molecules and Clusters*, 10, 135–144.
- Scully, M. O., & Drühl, K. (1982). Quantum eraser: A proposed photon correlation experiment concerning observation and “delayed choice” in quantum mechanics. *Physical Review A*, 25, 2208–2213.
- Scully, M. O., Englert, B.-G., & Schwinger, J. (1989). Spin coherence and Humpty-Dumpty. III. The effects of observation. *Physical Review A*, 40, 1775–1784.
- Scully, M. O., Englert, B.-G., & Walther, H. (1991). Quantum optical tests of complementarity. *Nature*, 351(6322), 111–116.
- Shalm, L. K., Meyer-Scott, E., Christensen, B. G., Bierhorst, P., Wayne, M. A., Stevens, M. J., Gerrits, T., Glancy, S., Hamel, D. R., Allman, M. S., Coakley, K. J., Dyer, S. D., Hodge, C., Lita, A. E., Verma, V. B., Lambrocco, C., Tortorici, E., Migdall, A. L., Zhang, Y., Kumor, D. R., Farr, W. H., Marsili, F., Shaw, M. D., Stern, J. A., Abellán, C., Amaya, W., Pruneri, V., Jennewein, T., Mitchell, M. W., Kwiat, P. G., Bienfang, J. C., Mirin, R. P., Knill, E., & Nam, S. W. (2015). Strong loophole-free test of local realism. *Physical Review Letters*, 115, 250402.
- Skotiniotis, M., Dür, W., & Sekatski, P. (2017). Macroscopic superpositions require tremendous measurement devices. *Quantum*, 1, 34.
- Smith, I. D., Krumm, M., Fiderer, L. J., Poulsen Nautrup, H., & Briegel, H. J. (2023). The min-entropy of classical-quantum combs for measurement-based applications. *Quantum*, 7, 1206.
- Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
- Terhal, B. M. (2015). Quantum error correction for quantum memories. *Reviews of Modern Physics*, 87, 307–346.
- Thomas, L. (1926). The motion of the spinning electron. *Nature*, 117(2945), 514.
- t’Hooft, G. (2016). *The cellular automaton interpretation of quantum mechanics*. Springer.
- Thorne, K. S. (2017). Nobel lecture delivered december 8, 2017: LIGO and gravitational waves III. Nobel Prize Outreach. www.nobelprize.org/prizes/physics/2017/thorne/lecture/.
- Uhlenbeck, G. E., & Goudsmit, S. (1925). Ersetzung der Hypothese vom unmechanischen Zwang durch eine Forderung bezüglich des inneren Verhaltens jedes einzelnen Elektrons. *Naturwissenschaften*, 47, 953–954.
- Uhlenbeck, G. E., & Goudsmit, S. (1926). Spinning electrons and the structure of spectra. *Nature*, 117(2938), 264–265.
- Van Helden, A. (1976). The importance of the transit of mercury of 1631. *Journal for the History of Astronomy*, 7(1), 1–10.
- Vandersypen, L. M. K., & Chuang, I. L. (2005). NMR techniques for quantum control and computation. *Reviews of Modern Physics*, 76, 1037–1069.
- Vidal, G. (2000). Entanglement monotones. *Journal of Modern Optics*, 47(2–3), 355–376.
- Von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik*. Julius Springer. English translation by R. T. Beyer: *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, Princeton, NJ 1955.
- Wang, J., Sciarrino, F., Laing, A., & Thompson, M. G. (2020). Integrated photonic quantum technologies. *Nature Photonics*, 14(5), 273–284.
- Werndl, C. (2011). On the observational equivalence of continuous-time deterministic and indeterministic descriptions. *European Journal for Philosophy of Science*, 1, 193–225.
- Werner, S. A., Colella, R., Overhauser, A. W., & Eagen, C. F. (1975). Observation of the phase shift of a neutron due to precession in a magnetic field. *Physical Review Letters*, 35, 1053–1055.
- Wheeler, J. A., & Zurek, W. H. (Eds.). (1984). *Quantum theory and measurement*. Princeton University Press.
- Wootters, W. K., & Zurek, W. H. (1982). A single quantum cannot be cloned. *Nature*, 299(5886), 802–803.
- Wu, X., Ma, Y., Yang, J., Chen, T., & Lin, J. (2019). On the security of TRNGs based on multiple ring oscillators. In *International conference on security and privacy in communication systems* (pp. 89–107). Springer.

- Xia, Z. (1992). The existence of noncollision singularities in Newtonian systems. *Annals of Mathematics*, 135(3), 411–468.
- Zhang, W., van Leent, T., Redeker, K., Garthoff, R., Schwonnek, R., Fertig, F., Eppelt, S., Rosenfeld, W., Scarani, V., Lim, C. C. W., & Weinfurter, H. (2022). A device-independent quantum key distribution system for distant users. *Nature*, 607(7920), 687–691.
- Żukowski, M., & Markiewicz, M. (2021). Physics and metaphysics of Wigner’s friends: Even performed premeasurements have no results. *Physical Review Letters*, 126, 130402.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter’s Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter’s Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 4

QBism: Integrating Agency and Quantum Mechanics



In this Chapter we build on our exposition of Quantum Mechanics (QM) in Chap. 3 to provide more detail on how agents can use QM as a tool, and on how we can integrate agency with quantum physics. We thereby elaborate specifically on the theme of Sect. 3.6, where we argued that agency and QM are virtually made for each other and constitute a package deal. We provide a concrete illustration of how QM is used in the general scientific process of building up knowledge and in the scientific quest of understanding the world around us, which includes ourselves and other agents. We also engage with the ongoing foundational debate about QM. Thereby we contribute to our book's overall aim to provide quantum-mechanical perspectives on the problem of agency.

From among the various interpretations of QM that are currently being discussed in quantum foundations, we single out QBism as closest in spirit to our approach, and this is why we mention QBism in this Chapter's title. For example, we agree with QBism in stressing the fundamental role of an agent as a user of QM, and in rejecting a large number of purported quantum paradoxes. We thus agree with the QBist diagnosis that many alleged conceptual problems of QM arise not from the nature of the theory, but from a failure to understand QM as a tool that agents can use. We do not subscribe to QBism wholesale, however—as we will point out, we reject the epistemological fundamentalism that is implied by QBism's exclusive focus on a single agent's experience as well as the QBist analysis of scientific rationality in terms of subjective betting odds. We also go beyond QBism in taking the notion of agency *itself* as a target for modelling rather than as an unresolved basic concept. That said, however, there are many aspects of QBism that we agree with wholeheartedly, and therefore it seems useful to stress these connections and to use QBism as our background for this Chapter. We provide a brief introduction to QBism in Sect. 4.2, but first we discuss the idea of integrating agency and quantum physics in general.

4.1 Integrating Agency and Quantum Physics

We distinguish three principally different ideas for how to integrate agency and quantum physics. (i) The first, head-on approach would be to aim at a full description of an agent as a physical system using only quantum mechanics. This would mean to assign to it some state vector $|\psi\rangle_{t_0}$ or density operator ρ_{t_0} at some given time t_0 and to use quantum dynamics to study the agent's time evolution, which could be entangled with the evolution of its environment. As a practical proposal for modelling a concrete agent, this approach certainly appears naïve, given the enormous dimension of the Hilbert space that would be involved. The real interest of this approach is, however, foundational: given that an agent—like any physical entity—is ultimately constituted by parts to which QM applies as a fundamental theory, what should prevent such modelling, at least in principle? (ii) The second, more limited form of integration would use QM for modelling only selected but important aspects of agency, for example, decision finding and learning, focusing on the potential role of specific quantum-mechanical resources such as indeterminism, coherence, and entanglement. Such an approach seems practically much more feasible. It could also contribute to current initiatives for modelling or even building quantum-enhanced learning agents. (iii) The third idea would focus not on the internal structure of an agent at all, but on how an agent can use quantum theory in practice, as a tool for getting around in the world and, especially, in the lab. This approach can be viewed as a domain-specific investigation into scientific agency, considering first those areas of science in which quantum mechanics figures explicitly in experimenting. It might, of course, generalise from there.

We comment briefly on all three approaches for integrating agency and QM in order to show how they relate to the discussion of this Chapter, and to QBism in particular.

(i). We are critical of the first idea, even in its foundational use, for fundamental reasons that we already discussed in the previous two Chapters. Of course, real agents are physical entities whose ultimate constituents are molecules, atoms, and elementary particles—we agree with that. But we also argued that agents must be viewed as ontologically emergent entities that organise their matter as material (see Sect. 2.2.3). It is at the level of the organised material, not at the level of the underlying matter, that agency must be described, otherwise it cannot be captured in its organisational specificity at all. This implies in particular that agents cannot be described as isolated quantum systems. Consequently, agents cannot be in quantum-mechanical superpositions of macroscopically distinct states—matter can be in superposition, but agents organising their matter as material cannot. Thought experiments starting from the assumption of agents being in quantum superpositions therefore appear misguided. Cats being agents, this means that we reject the thought experiment of Schrödinger's cat (thereby agreeing with

Schrödinger’s own intention in cooking up the story).¹ It also means that we are critical of Wigner’s thought experiment (a variation on Schrödinger’s cat) in which Wigner’s friend conducts a quantum experiment in an isolated laboratory while Wigner himself describes that laboratory from the outside. Wigner held that there was a paradoxical tension between the friend’s view, who would be sure of having registered a unique outcome, and his own, outside observer’s view: He would have to describe his friend’s state as a superposition rather than as a mixture of the different possible measurement outcomes. Only upon receiving information about the friend’s measurement outcome would Wigner himself be in a position to update (or “collapse”) his state ascription. We agree that there is a tension, but we reject Wigner’s reasoning: If his friend has really registered an outcome, then an irreversible process has really occurred, and Wigner’s uncertainty about the outcome must be described in terms of a mixed state rather than a superposition. If, on the other hand, his friend has not yet registered an outcome, then we are discussing just a pre-measurement rather than a measurement, and no conclusions can be drawn.² In a similar vein, our view of agents as ontologically emergent individuals gives us reason to reject self-referential scenarios in which “quantum mechanics is used to describe its own use” (Frauchiger & Renner 2018).³ In our general assessment of the first idea for integrating agency and QM, we are thus fully in line with QBism (see DeBroda et al. 2020), even though we differ from QBism in grounding our analysis in an ontological understanding of irreversibility.

(ii). We are convinced of the value of the second idea of a more limited and more specific integration of agency and QM in which quantum resources are seen as enabling or enhancing agency. For a simple example, consider an agent who controls individual nuclear spins in its environment for quantum sensing of weak magnetic fields. We also already argued for the positive role of QM’s basic indeterminism as enabling ontological emergence in Sect. 3.6 above. In this and the following Chapters we will broaden this approach, looking at central aspects of agency such as decision finding, memory, and learning.⁴ We will investigate how these aspects of agency can be understood based on indeterminism as a resource, or can even be

¹ Schrödinger (1935) after all called his cat scenario a “ridiculous case” (the original German has “burlesk”, which means “comical in a crude way”).—Quite generally, producing a measurement outcome (which can be, but need not be, experienced by an agent) involves an amplification process, e.g., via a photodetector, which is irreversible. See Glauber (1986) for a theoretical model of an amplifier and its application to the Schrödinger cat problem, and see Sect. 3.3.2.3 for some general comments on irreversibility.

² Recall the slogan from the previous Chap. 3 (footnote 68): Unperformed measurements have no outcomes; and even performed pre-measurements have no outcomes either (Żukowski & Markiewicz 2021).

³ Generally, the notion of agency that is used in such thought experiments is very thin. Agents are typically reduced to carriers of a state of belief or knowledge. They may also have some capacity of logical inference, but neither their having beliefs nor their reasoning capacity is normally modelled explicitly with a view to its physical basis. See Sect. 5.6 for some concrete, model-focused remarks.

⁴ For a specific proposal in this direction, which we will address in more detail in Sect. 5.6.4, see Flamini et al. (2024). See also Jerbi et al. (2023).

modelled in terms of quantum processes, which connects with the idea of quantum machine learning (Dunjko & Briegel 2018). This line of inquiry motivates the idea of modelling natural as well as artificial agency, which will be central to the discussion of Projective Simulation in Chap. 5. Prior to that, we will discuss agents as modellers and agency modelling in Sect. 4.4–4.5 below, keeping the discussion mostly at the level of classical physics plus indeterminism. The idea of agency modelling goes beyond the current general state of discussion of QBism, where the notion of an agent is an unresolved primitive.⁵

(iii). We are also convinced of the value of the third idea of integrating agency and QM. That approach aims at a formal description of scientific agency, which includes learning and experimenting, specifically in application to systems that fall within the scope of QM. It thereby promises insights into the notion of a scientific theory as a tool that agents can use. QBism's research program is primarily based on this approach. Experience gained through quantum experiments stands out among other forms of experience due to its fundamentally participatory nature, which explains the strong focus that QBism places on quantum measurement and on agents' quantum state updating. The goal is to develop QM as a general theory of how we, and possibly even artificial agents, can learn about the world through experience. This may then prepare the ground for the use of artificial agents in "self-driving" quantum laboratories and in automated scientific discovery (see our discussion in Sect. 5.4).

An integration of agency and QM according to this third idea can in principle be combined with approaches for resolving the notion of an agent both along the lines of ideas (i) or (ii). While we already voiced criticism about a "fully quantum" view of an agent as presupposed by idea (i), we are open towards a combination of (natural or artificial) scientific agency with a quantum view of the world. Different notions of quantum agents have already been studied,⁶ and the idea of an agent having quantum-coherent access to its environment is fascinating and challenging.⁷ Our main modelling effort will, however, be geared towards a combination of approaches (ii) and (iii) in their respective contexts, through which we hope to bring agency modelling and QBism together in a way that goes beyond the limits of the current QBist discussion.

⁵ See also the critical discussion in French (2023a, Ch. 8) and in French (2023b). For a QBist development that does align with the mentioned approach of agency modelling, see Pienaar (2020).

⁶ See, for example, Paparo et al. (2014), Dunjko et al. (2016), and Dunjko and Briegel (2018) and, more recently, Kewming et al. (2021) and Elliott et al. (2022). See also our discussion in Sect. 5.6.

⁷ A first hurdle to overcome is to provide an adequate formal description of a quantum agent: What does learning mean for such an agent, and how can the notion of a learning history be understood? These questions are at the forefront of current research. See Dunjko et al. (2016) for a general, abstract proposal.

4.2 Introducing QBism

QBism is a recent and evolving research program in the foundations of quantum mechanics. It takes inspiration from a number of earlier attempts to understand QM that stress the subjective nature of quantum state updates and the formal similarity between such updates and Bayesian learning. QBism thereby incorporates lessons from several of the pioneers of QM. Fuchs (2017a) has said that “[w]ithout Bohr, QBism would be nothing”, and Mermin (2017) traces out connections to Schrödinger, stressing the QBist idea of quantum mechanics as an agent’s “user’s manual”. Branching off from other forms of “quantum Bayesianism”, QBism emphasises the fundamental role of agents and their experiences.⁸ There is a lively debate on the core tenets of QBism, fueled in part by perceived misunderstandings of the approach by its critics, which has led to an extensive FAQ list (DeBroda & Stacey 2018) as well as a number of articles by the main proponents of QBism in which core aspects of QBism are clarified (see, e.g., Fuchs 2010; Fuchs et al. 2014; Mermin 2017; Stacey 2019; Pienaar 2020; DeBroda et al. 2021; Fuchs et al. 2022; DeBroda et al. 2024). Recent critical evaluations include Earman (2019), Dieks (2022), and French (2023b).

From our agency-centric perspective, QBism is interesting because it takes very seriously the idea that quantum mechanics is a *tool* that agents can use to get around in the world. QBism takes the subjective nature of an agent’s quantum state update to the extreme, so that QM is understood as a single-user theory. Consequently, QBism views an agent’s individual experiences as the sole authoritative basis for that agent’s dealings with the world, e.g., via experiments. Bayesian learning in the framework of QM becomes *the* tool for agents to organize their experiences. In order to use this framework, agents have to assign probabilities to their expectations. In line with its general single-agent-centred approach, QBism subscribes to a strictly subjective, personalist interpretation of probabilities, inspired by De Finetti’s (originally all-caps) pronouncement that “probability does not exist” (De Finetti 2017, xv): Probabilities are not forced upon agents as objective features of their environment; rather they are tools that agents construct and use to get around. The standard axioms of probability theory are justified pragmatically via a so-called “Dutch book” argument: agents who violate the probability axioms can be exploited via a certain system of bets called a Dutch book (see Sect. 4.2.2 below for discussion). Conforming to probability theory, an agent gaining new experiences should update her personal probabilities in a Bayesian fashion, based on the prior probabilities she had assigned to possible new experiences. Importantly, even probability-1 or probability-0 assignments are fundamentally still seen as probability assignments, which precludes their use in an EPR-like “criterion of reality”.⁹

⁸ The label “QBism” was first used in 2009 (Stacey 2019). See Healey (2023) for a discussion in the context of other Bayesian frameworks of QM; see Pienaar (2021) for a comparison with Rovelli’s relational QM.

The quantum state plays a central role for QBist updates, providing a manual for probabilistic expectations. In foundational terminology, its status is neither ψ -ontic nor ψ -epistemic, but pragmatic: it is a mathematical tool for agents to adjust their expectations about future experiences in future experiments. The Born rule, which provides the probabilities that regulate an agent's expectations, is seen as “an empirically motivated addition to probability theory—a normative requirement of quantum Bayesian coherence [...] that supplements the usual coherence conditions on degrees of belief required to avoid a Dutch book” (Healey 2023). In this way, QM becomes part of scientific rationality understood in broadly Bayesian terms: agents act under the normative requirement to have probabilistic expectations that conform to QM's Born-rule probabilities.

This view of the Born rule is connected with an ambitious technical research agenda in QBism, namely, the full probabilification of QM. The main lines of this program are described by Fuchs and Schack (2013). It rests on the assumption that, for any Hilbert space dimension, one can find a distinguished set of observables which form a special type of POVM which is *symmetric* and *informationally complete* (SIC). Recall that a POVM, as described in Sect. 3.3.2.2, is specified by a set of observables $\{\hat{O}_i\}$ that are positive semi-definite and sum up to the identity, $\sum_i \hat{O}_i = \hat{1}$. A POVM describes a generalized measurement with outcomes $i \in I$ and corresponding probabilities $\text{Pr}(i) = \text{tr}(\hat{O}_i \rho)$. A SIC-POVM is a POVM with the additional properties that, for a Hilbert space with dimension d , the POVM has precisely d^2 elements and the measurement observables are projectors onto pure states, with a constant quantum mechanical overlap between each pair of projectors. Mathematically, a SIC-POVM is specified by the relations

$$\hat{O}_j = \frac{1}{d} \Pi_j \equiv \frac{1}{d} |\psi_j\rangle\langle\psi_j| \quad (4.1)$$

with

$$\text{tr}(\Pi_j \Pi_k) = |\langle\psi_j|\psi_k\rangle|^2 = \frac{d\delta_{jk} + 1}{d + 1}, \quad (4.2)$$

for $j, k \in \{1, \dots, d^2\}$.¹⁰ From the last property (Eq. (4.2)) it follows that the d^2 projectors $\Pi_j = |\psi_j\rangle\langle\psi_j|$ are linearly independent and thereby form an operator basis for all hermitean operators acting on the Hilbert space of dimension d (Renes et al. 2004). As a consequence, if a SIC-POVM exists in a given Hilbert space of dimension d , any state ρ in that Hilbert space can be expressed as a linear

⁹ We briefly mentioned that criterion when discussing hidden variables in Sect. 3.4.1. See Fuchs and Schack (2015) for a discussion in the QBism context.

¹⁰ The property of constant mutual overlap between different states $|\psi_j\rangle$ in Eq. (4.2) is reminiscent of the complementarity relations of Eq. (3.92). Note, however, that for $d = 2$ the SIC-POVM contains $d^2 = 4$ elements.

combination of the POVM's elements,

$$\rho = \sum_{j=1}^{d^2} \left((d+1) \Pr(j) - \frac{1}{d} \right) \Pi_j. \quad (4.3)$$

The significance of this equation lies in the fact that the quantum state can be written entirely in terms of generalised measurement operators $\hat{O}_j$ and the probabilities $\Pr(j) = \text{tr}(\rho \hat{O}_j)$ associated with the measurement outcomes.

Building on this observation, and on other properties of SIC-POVMs, Fuchs and Schack (2013) show how one can write the Born rule as an equation that only contains probabilities, without the need of referring to complex probability amplitudes and phases of quantum states, as in the standard formulation in QM. In doing so, they understand the Born rule “as a relation between probabilities, rather than a *setter* of probabilities” (Fuchs & Schack 2013, 1709).

As of today, there is no proof that SIC-POVMs do exist in all dimensions, even though SIC-POVMs have been found for many dimensions. For a recent survey of the state of play in this field see Fuchs et al. (2017). We do not expand on this aspect of QBism as it is tangential to our interest in this book. We agree, however, that it holds great promise for a deeper structural and mathematical underpinning of the QBist program and its relations to other fields, for example the recently discovered connections of the SIC problem to open questions in algebraic number theory (Appleby et al. 2017; Bengtsson 2020).

4.2.1 *QBism, Quantum Paradoxes, and Common Criticisms*

One of the clear benefits of QBism is that it resolves, from the outset, several quantum paradoxes that most other interpretations struggle with. These include the “measurement problem”, the “Wigner’s friend paradox” (see, e.g., Fuchs 2010) and its recent extensions (DeBroda et al. 2020), and the problem of non-locality (Fuchs et al. 2014)—we agree with the basic thrust of the QBist approach here.¹¹ Another benefit is its conceptual simplicity and the consistent integration of quantum mechanics with Bayesian probability theory, which seems attractive. Still, the strong focus on agents’ personal experiences makes QBism unpalatable for many physicists, as witnessed by the intense debate surrounding the framework, both in scholarly works and in blogs. Shouldn’t a fundamental theory of physics focus on the real world and describe reality as it is independently of any agents, rather than focusing on agents’ belief-updating strategies? Such worries are enhanced

¹¹ See Sect. 3.3.2.2 and our discussion of option (i) in Sect. 4.1 above. As we pointed out in Sect. 3.3.2.3 and as will discuss in more detail below, however, we are not afraid of ontology and of speaking of indeterministically occurring irreversible events. Our own position is thus closer to Englert (2013) than to the letter of QBism.

by QBism's insistence on a strictly personal interpretation of probability, and on viewing quantum mechanics as a single-user theory.

At this point, two types of criticism are often levelled against QBism—unfairly, as we believe. First, it is not uncommon to find QBism associated with solipsism, which everybody agrees is a sad and lonely affair. The charge of solipsism, however, appears misguided, as QBism's pragmatic stance assumes the existence of the world from the outset: The very meaning of an experiment implicitly presupposes the existence of a world that reacts to the experimenter's intervention.¹² Second, a further charge is that QBism subscribes to anti-realism in the form of a mere instrumentalism. Instrumentalism has, however, two readings that need to be kept apart. It is true that QBism stresses the role of theories—QM or others—as tools and thus, as instruments. In general, there is nothing wrong with such a pragmatic view: theories are there because they are useful for us, enhancing the range of our agency in manifold ways. There is, however, also a pejorative use of “instrumentalism”, which is applied critically against those who understand theories as fundamentally and exclusively at the service of “saving the phenomena” and without any value of their own. The main example anchoring this pejorative use is Ptolemaic astronomy, which accounted for observations—thus saving the phenomena—via a baroque multiplication of epicycles.¹³ That approach may actually not be so far from current trends of “big data-only science”. But to view science as merely saving the phenomena greatly underestimates the theoretical (e.g., explanatory) and practical (action-guiding) value of theories for science. At any rate, the QBist interpretation of theories as tools is not meant to downplay their value at all. It is simply a fact that we use theories such as QM as tools to good effect: for basic and explorative research, to design and analyse new experiments, and to investigate how matter behaves at the atomic scale. It should be uncontroversial that theories are, and have always been, instruments in that sense.

4.2.2 *Going Beyond QBism*

Even though QBism can be defended against the charge of solipsism and against the pejorative accusation of instrumentalism, we are critical of some aspects of QBism's current philosophical outlook. Given that QBism is an evolving research program, we do not mean these criticisms as a refutation of the approach, but rather as an indication of relevant areas of research in its development. We mention two core

¹² For a general defense of Quantum Bayesianism against solipsism, see Timpson (2008). For a discussion that addresses QBism specifically, see Glick (2021).

¹³ It is easy, but somewhat anachronistic to dismiss the strategy of “merely saving the phenomena” as bad science not interested in theory-building. In the medieval historical context one needs to take into account the powerful role of the church that, broadly speaking, was hostile towards speculations about deeper causes of the phenomena, which made theory-building politically dangerous. This situation has changed.

issues, both connected with our broader understanding of (scientific) agency and the ontologically emergent nature of agents: epistemological fundamentalism and probabilism based on propositionalism. We hold that on both accounts, an adequate understanding of agency requires us to go beyond QBism.

Epistemological Fundamentalism One of the most critical aspects of the epistemological side of QBism, in our view, is its complete reliance on a further unresolved notion of an agent's "own direct internal awareness of her own private experience" (Fuchs et al. 2014, 750). The danger we see here is that such experiences, while admittedly leading to updated expectations, are themselves assumed to be incorrigible, committing QBism to epistemological fundamentalism. It is of course attractive to try to build up a view of agents' scientific dealings with the world from solid and unshakeable foundations, but the history of such attempts suggests that incorrigibility cannot be defended.

Famously, Descartes tried to find firm epistemological foundations in his *Meditationes*, using his approach of methodological scepticism. Following his *cogito* argument, he came up with the idea that veridical, trustworthy experiences should wear their veridicality on their sleeves, being "clear and distinct".¹⁴ But could one not be mistaken in discerning this mark? Descartes admitted that an agent alone could not find an adequate guarantee in her personal experiences. Even for the basic defense of his approach he had to resort to a theological warrant via divine benevolence. And the possibility of agents' being confused about the clarity and distinctness of their experiences still remained a threat that could not be satisfactorily resolved.

In the twentieth century, the Vienna Circle logical empiricists made a fresh attempt at building up scientific knowledge from solid epistemological foundations. Several approaches were tried, as documented in the famous 1930s debate about protocol sentences (see Uebel 2007). But the project had to be abandoned—neither a foundation based on sense data nor the idea of protocol sentences in the language of physics could do the job. Any agent's experiences—including those concerning experimental outcomes—are fallible and may have to be corrected in the larger individual and social context of later actions and experiences. Both the world and other agents are necessary parts of these dynamics; they cannot be crossed out. Experience reaches into the world, and even the stability of the content of an agent's remembered experiences cannot be justified without the contribution of other agents: A single agent alone is incapable of grounding normativity, but meaningful experiences presuppose a distinction between getting it right and getting it wrong.¹⁵ Especially in science, intersubjective stability of beliefs is key, and this

¹⁴ See Cottingham et al. (1984, Vol. 2, 24) for the passage in the standard English translation. For the original, see Adam and Tannery (1904, Vol. 7, 35).

¹⁵ Basically this is the upshot of Wittgenstein's much-discussed argument against the possibility of a private language, which he developed in the 1930s after his intense discussions with members of the Vienna Circle. Wittgenstein had himself previously attempted to analyse the elementary

presupposes an acknowledgement of other agents' views of the world as equally important as any focal agent's view (see Oreskes 2019).

Acknowledging fallibilism and the fundamental importance of intersubjectivity does not commit one to a full ontologisation of one's theories. Specifically, viewing QM as an intersubjective endeavour does not require one to view the quantum state as "out there" in the world—that state can still be viewed as (part of) a tool that individual agents use from their own perspective to get along, in full agreement with QBism. But that tool reaches into the world, together with similar tools that other agents can use to reach into that same world. And in science, one joins forces with these other agents to understand how it is that the world pushes back, and to jointly develop adequate concepts, models, and theories.¹⁶ In our view, a broader notion of learning can help to more properly anchor these important aspects of (scientific) agency. This ties in with our second point.

Propositionalism and Probabilism We fully agree with QBism that there is a useful parallel to be drawn between an agent's updating a quantum state ascription and the Bayesian formalism of belief updates. Bayesian updates take place in a given space of possibilities that can be formally specified via structured propositions. Once an agent has identified a quantum system and a measurement, or an observable, the Born rule can be usefully understood as a normative demand on that agent to set her personal probabilistic expectations in accord with the rule's predictions (for example, in the simplest case, to use Eq. (3.43), $\Pr(a_i) = |\langle a_i | \psi \rangle|^2$, or, more generally, $\Pr(i) = \text{tr}(\Pi_i \rho)$, with Π_i as defined in Eqs. (4.1) and (4.2)).¹⁷

As we will indicate below through a concrete example (Sect. 4.4.1), however, not all learning in the context of QM can be neatly captured in this scheme. In many interesting cases, a probability space in which an update should take place cannot be identified ahead of the experiences that will teach an agent something. QBism, however, appears to presuppose this, and is thereby committed to two philosophical theses about agents and their beliefs that we find questionable: propositionalism and probabilism. Both assumptions, while warranted in the mentioned special case of quantum state updating, seem dubious as general assumptions about how agents learn and structure their experiences and expectations.

Propositionalism is the thesis that all of an agent's mental states, including knowledge and belief, can be captured in terms of propositions, which are expressed via phrases of the form "that p ". Thus, Mary may believe *that it is raining*, or know *that $2 + 2 = 4$* , or learn *that John is in Paris*. In philosophy there is an intense debate on whether this scheme really generalises to all mental states (see, e.g.,

propositions presupposed in his *Tractatus* in terms of individual sense data, but had then abandoned the idea as hopeless. See Candlish and Wrisley (2019) for an overview.

¹⁶ French (2023b) similarly stresses the importance of such a social setting in his investigation of how QBism aligns with phenomenology, which we will take up in Chap. 6.

¹⁷ The agent may also use the QBist version of the Born rule (DeBroda & Love 2022); this makes no difference for what follows.

Montague 2007; Felappi 2023). It will not be necessary here to enter into the details of that discussion. It suffices to note that in the context of agency and learning, an important recalcitrant notion is that of skill, or know-how. Mary may not just know *that* $2 + 2 = 4$, but also *how to ride a bike*, or *her way around London*. Acquiring such skills is an important form of learning that is rooted in embodied social interaction, and the resulting know-how appears ill suited to a reformulation in terms of “knowing that”.¹⁸ It also seems reasonable to assume that agents, in their individual development, acquire lots of know-how before they advance to the stage of knowing-that, if they do at all.¹⁹ The point is not that we cannot describe an agent’s quantum state-ascriptions in terms of propositional beliefs, but that it seems wrong to assume that agential learning is always an explicitly propositional affair.

Apart from propositionalism, QBism subscribes to probabilism, which is the thesis that an agent’s (propositional) beliefs should be graded in terms of their strength, and that such gradation must rationally come in the form of probabilities assigned to the propositions. Probabilism about beliefs in fact presupposes propositionalism about beliefs because the probabilistic norms that govern the strength of beliefs (e.g., that for mutually exclusive A and B , $\Pr(A \vee B) = \Pr(A) + \Pr(B)$) presuppose the applicability of propositional connectives such as conjunction ($\wedge$), disjunction ($\vee$), and negation ($\neg$). QBists employ a philosophically well entrenched strategy of arguing for probabilism, so-called Dutch book coherence. Very briefly, the assumption is that an agent’s assigning probability p to proposition A can be analysed in terms of the agent’s betting behaviour: such an agent buys or sells a bet that pays out 1 \$ if A is true, if and only if the price for that bet is p \$. On this analysis of what the strength of beliefs comes down to, it can be shown that agents who violate the axioms of probability theory are susceptible to being exploited via a so-called Dutch book, which instructs an adversary (e.g., a casino) to buy and sell bets in such a way that the agent incurs a sure loss. For example, if our agent sets her personal probabilities for the mutually exclusive propositions A and B to be $\Pr(A) = p$ and $\Pr(B) = q$, but assigns probability $\Pr(A \vee B) = r < p + q$, the adversary can sell two bets to the agent, one that pays 1 \$ if A is true, and one that pays 1 \$ if B is true, at the respective prices of p \$ and q \$, and buy a bet from the agent such that the agent will have to pay 1 \$ if $A \vee B$ is true, at the price of r \$. In all possible circumstances, the agent ends up with a net loss. After buying and selling the bets, the agent has paid $(p + q)$ \$ and received r \$, which totals $(-p - q + r)$ \$ < 0 ; the adversary (casino) has $(p + q - r)$ \$ > 0 . No matter how A and B turn out, the agent and the adversary will not gain or lose anything in addition. If both A and B turn out false, none of the bets becomes due, while if either A or B turns

¹⁸ For example, it has been shown that people have egregiously mistaken propositional beliefs about how cycling works, even if they know how to ride a bike (Lawson 2006).

¹⁹ One influential position in the philosophical discussion here is due to Ryle (1949), who is famous for his criticism of what he called intellectualism, and for the thesis that (in a certain technical sense) knowing-how is conceptually prior to knowing-that. (See Bäckström and Gustafsson 2017, for discussion.) See Chap. 6 for a phenomenological take on the primacy of everyday knowing-how.

out true (which guarantees that the other turns out false, as *A* and *B* were assumed to be mutually exclusive), the agent and the adversary both have to pay 1 \$ to the other party, netting zero. The agent thus sticks with the loss made when buying and selling the bets.

The Dutch book argument for probabilism may seem convincing, but it is philosophically disputed (see Vineberg 2022). There are other arguments for probabilism based on other considerations, for example, so-called accuracy arguments (Pettigrew 2016), that may fare better in some respects. But each approach to justifying probabilism is contentious, and it turns out that fine details in the exposition matter a lot (Hájek 2008). From an agency perspective, we have already voiced objections to propositionalism, so that probabilism across the board is not an option anyway. There are forms of agential learning that cannot be covered by probabilistic updates, simply because the agents gain non-propositional knowledge (e.g., motor skills). But even in the case of propositional beliefs, we are sceptical of the assumption that agents attach probabilities to all of them, and we are especially sceptical of the analysis of such probabilities in terms of betting. Agents sometimes try out things without having any expectations either way, much less quantifiable expectations, and they often have better things to do than bet. Furthermore, they may hold beliefs on whose truth or falsity they will simply never bet.²⁰ Additionally, it seems just wrong to assume that agents are irrational when they do not accept a fair—or even a favourable—bet.²¹ Again, this is not to say that we disagree with the QBist analysis of quantum state updates. But as a general story about agency, propositionalism and probabilism cannot be upheld.

Going Beyond QBism The two criticisms of QBism that we just formulated are connected: they are both directed against the QBist assumption of starting from very thin and further unresolved notions of agency and of the world. QBism puts an agent's personal experiences centre stage. Such experiences (*E*) necessarily belong to a specific agent (*A*), and they also pertain to the world (*W*). In the current formulation of QBism, however, both the agent and the world have a merely systematic role: They appear as reference points for the experiences, but they are themselves not resolved to have internal structure (see Fig. 4.1). We hope to be able to go beyond the current state of QBism at this point.

We agree that any theory must presuppose some notions that it cannot resolve any further (see Sect. 3.1). But our project is to understand agency as a real feature

²⁰ Relatedly, Milgram (2005, Ch. 10) argues that there must be something seriously wrong with agents whose actions are fully determined by a utility function that allows them to weigh *all* possible courses of action against each other. Practical decision making is not just an optimisation task; agents may be committed to values that they express and strengthen through their decisions. See Kane (1998, 2024) for pertinent remarks on what he has termed self-forming actions.

²¹ On 8 November 2022, a person in the U.S. won the powerball lottery jackpot of 2.04 billion US\$. A ticket cost 2\$, and winning odds for the jackpot were 1:292 million. That bet was thus highly favourable, even taking taxes etc. into account. It seems a stretch, however, to accuse everybody who did not play of irrationality. (Strictly speaking, even the person who won could be criticised for not having bought more tickets to increase his chances.)

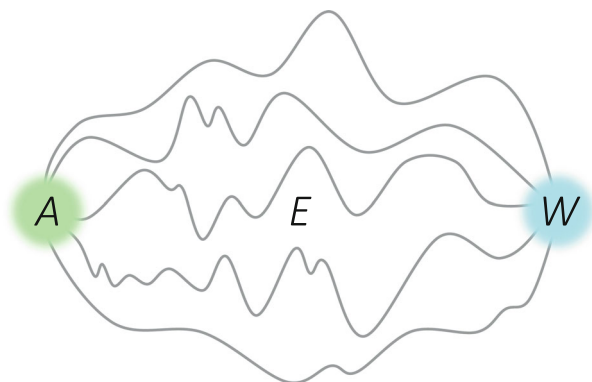


Fig. 4.1 The framework of current QBism: Agent (A) and world (W) are unresolved reference points for the agent's experiences (E)

of our physical world, and so we have to resolve the notions of agent and world beyond their role as merely limiting notions as assumed in QBism. We need to allow for agency *itself* to be the subject of physical theories such as QM. As we said in Sect. 4.1, this does not mean that we should approach an agent as a quantum system, but we will try to link specific aspects of agency to QM. In our view, it should be possible to model agency as a feature of our physical world, and to show how agents participate in what happens in the real world.

4.3 Participatory Realism

QBists subscribe to “participatory realism”, which holds promise for breaking the deadlock between scientific realism vs. anti-realism, especially as concerns QM (Fuchs 2017b). But there is work to do to develop the view in more detail.

4.3.1 *Realism vs. Anti-Realism in the Philosophy of Science*

One of the big questions in the philosophy of science concerns the status of scientific theories and the entities that are referenced in them. Should the claims of accepted scientific theories be taken literally, as statements about what the world is like, or should the theories rather be viewed as helpful tools that have no ontological import? These two options correspond, broadly speaking, to the positions of scientific realism and anti-realism. The former option seems to be the standard view of the working scientist: when one researches quarks, genes, or El Niño, it is customary and useful to adopt a naïve realist stance according to which these entities are simply out there in the world to be studied. But the latter option also has much going for

it, especially when one reflects on the history of science as the empirical basis that should be used to adjudicate between the two views (Laudan 1981): it is specifically the ontological claims of scientific theories that tend to be superseded in the history of science. Insofar as there is stability across theory change at all, it is due to the structures posited by successive theories rather than their claims about what the world is like (Worrall 1989; Ladyman 2023). Both realism and anti-realism receive additional support through further arguments. In its general form, the debate remains unresolved, at least to the extent that both positions are still advocated and defended by serious and cogent arguments.

In QM, the realism/anti-realism issue is complicated by the fact that it is hard to say what naïve—or even not-so-naïve—realism should amount to. Perhaps linked to that difficulty, there are attempts at having one’s cake and eating it, too, which leads to conceptual muddles. The realism debate in QM mostly revolves around the interpretation of the wave function ψ , through which a quantum state can be expressed in terms of complex-valued amplitudes associated with, e.g., points in space (see Sect. 3.3.2.2). The realist approach is to try to understand ψ in analogy to a physical field (e.g., the electromagnetic field), and to provide a reading of the wave function as the faithful representation of something that is out there in the world independently of our interactions with it. An example of such a reading is given by Bohmian mechanics, in which the world is represented as a system of particles in space whose motion is guided by some real structure, the pilot wave, which can be directly derived from ψ . This picture, as we already remarked, amounts to agency scepticism: agents and their actions (e.g., their measurements) are all part of the big show and have no fundamentally different status from the rest of the world—there are, simply, no agents as ontologically emergent entities (Bohm & Hiley 1993). While such a fully representational reading may convey a consistent metaphysical picture, its usefulness and operational value in active research and experiment is limited. A similar assessment seems appropriate for other hidden variables alternatives to QM, and also for the various “many worlds” approaches to understanding “the wave function of the universe”.²²

Anti-realism about QM can also take many forms. One extreme is a mere instrumentalism whose only slogan is “to save the phenomena”. We already concluded that viewing QBism as committed to anti-realism of that type seems unwarranted (see Sect. 4.2.1). The somewhat ill-defined “Copenhagen interpretation” of QM—the way of using and developing that theory initiated by Bohr and his colleagues at Copenhagen—is often associated with anti-realism, sometimes linked to pejorative slogans such as “QM works, shut up and calculate”. But a rejection of (naïve) realism about QM can take other, more interesting forms. For example,

²² We are aware of the fact that this assessment will be controversial. The main point for us is that the notion of agency in these approaches is either explicitly excluded (e.g., in superdeterministic theories such as t’Hooft 2016) or seriously underresearched (see, e.g., Vaidman 2021). We hold realism about agency to be much more important than any realist view about specific scientific theories, and so we stick to our project of developing our account of real scientific agency in QM without further addressing realism about the wave function.

“Copenhagen” can also be viewed, more adequately, as the attempt of providing an *interactive* view of QM in which the experimental probing of nature by scientists is taken as centrally important. Historically, the scientific agents were understood as mere “observers”,²³ but an interactive view of QM can go beyond this limitation. In fact, QBism can be viewed as a useful radicalisation of the Copenhagen approach in this respect: the observer is replaced by an agent, and the agent has an absolutely central systematic place in the theory.²⁴

4.3.2 *QBism and Participatory Realism*

QBists themselves, as we said, describe their view as “participatory realism”, which aligns well with an interactive view of QM: the world is really out there, and QM provides exceptionally precise and useful tools that agents can use for getting around in it. A representational view of QM is neither necessary nor fruitful, leading to conceptual problems mostly unrelated to scientific practice. An interactive view of QM that stresses the theory’s participatory aspects avoids these problems from the outset and thus seems much more attractive.

So far, however, QBism’s participatory realism appears to be mostly a label for a research program. Many interesting ideas and intriguing connections to the philosophical traditions of pragmatism and phenomenology have been pointed out, but neither the participants nor the reality of participatory realism have been much discussed (see French 2023b). At this point, we also hope to be able to go beyond the current state of the QBist discussion, with respect to all the three components of agency, world, and their interaction. We will discuss how agents participate in worldly happenings through modelling them, and how we can model agency itself as a phenomenon that really exists in the world. The meaning of participation or interaction will also need to be clarified. In an interactive approach, the world is not simply the subject of our experience, but the subject and partner of our interactions. As embodied agents, we interact with the world through our bodies, but also through *tools*, which may be practical or theoretical (see Sect. 6.6.2). A tool points both to the world and to the agent, and when one employs tools, one experiences the world as resisting. This resistance provides constraints to our agency, which we as scientific agents may try to overcome in turn by devising new, improved model- or theory-guided ways of interacting with the world.

²³ According to Bohr (1934), a classical, macroscopic reference frame is necessary in order to ground even the most basic notions such as that of an experiment. Later on, the role of classical physics and the role of an observer in the description of measurement processes have been much discussed (see, e.g., Wheeler and Zurek 1984).

²⁴ In QBism, measurement is not seen as a physical process, but as an experience-generating act of a given agent (Fuchs et al. 2014). QBism hereby rejects the idea of objective mechanisms that govern the world. Our aim of resolving the agent as a physical system means that we have to go beyond QBism at this point.

On an interactive understanding of QM, the wave function ψ itself is a good example of a theoretical tool. As such, ψ is about the world, not in a representational sense, but pragmatically, in the sense of providing guidance for action and anchoring expectations for experimental interactions. It is, for example, a quantity that we can use to predict the outcome of a scattering experiment, where $\psi(\theta, \varphi)$ together with the Born rule is used to determine the scattering cross section for given angles θ and φ . Another example is the determination of the possible energies that a system can attain, where we use ψ indirectly in order to compute the eigenvalues of the Hamiltonian. Furthermore, the wave function can also guide our intuition in thinking about various quantum phenomena—such as chemical binding, the notion of molecular orbitals, super-fluidity, nuclear decay, or laser light—and inspire the design of new experiments, for example in quantum optics or in condensed matter physics. It is a tool with many applications in the investigation of light and matter at the quantum scale.

In the practice of basic research in physics, we see human explorative agency at work. Philosophical questions regarding the correct interpretation of ψ are often moot. But researchers must have a good understanding of the systems and phenomena they are dealing with, especially if they try to predict the outcomes of previously unperformed experiments. Prediction is certainly not the only goal of basic research, but successful prediction of novel phenomena provides evidence for understanding, which often includes insights into mechanisms. And the mechanisms underlying a given phenomenon are naturally taken to be objective properties of the world.

Gaining scientific understanding is an interactive and necessarily intersubjective learning process. Participatory realism will need to develop this aspect of learning in a way that is broader than QBism's current approach that centres on Bayesian updating. That scheme is too narrow especially in the most interesting situations in research, when we try to understand a new phenomenon. Our aim will be to work out a more adequate interactive view of understanding and learning in science by taking into account the interplay of the development of models, theories, and experiments in basic research.²⁵ The notion of modelling will be central for our account. In fact, we believe that it is useful quite generally to view scientific agency as modelling.

4.4 Scientific Agency as Modelling

The notion of a model plays an important role in science as well as in philosophy of science. There are many different types of models, with different uses, so that there is no simple formal definition of what a model is (Frigg & Hartmann 2025). From our agency-centred perspective, it is most interesting to look not at models as

²⁵ In line with QBism, we expect that concepts from phenomenology will be helpful to describe the pertinent notion of tool in more detail. See Sects. 6.2.3 and 6.6.2 for a discussion of tools and models from a phenomenological point of view.

finished products of some endeavour, but at the process of modelling itself, which plays a crucial role in understanding nature and in learning about the world.

Learning as modelling can take many forms, depending on how difficult and challenging the setting is for the agent. We can distinguish three broad levels of difficulty. The simplest case is captured by the QBist ideal of learning via Bayesian updating, where an agent's model of a system takes the form of a probability distribution over possible states. This is often appropriate when the system under consideration is sufficiently well understood and under control. As a concrete example, consider a collection of trapped ions whose internal energy states constitute the register of a quantum computer. To characterise the quantum state and to demonstrate, e.g., that entanglement of a certain type was created, a variety of statistical methods have been developed.²⁶ For a full characterisation of the state of quantum system, quantum tomographic techniques (Paris & Řeháček 2004; Gross et al. 2010) are the gold standard.²⁷ While Bayesian methods are not typically used in this context, they are a viable option (Lukens et al. 2020). One can therefore agree that the QBist ideal of learning via updates in a given probability space does capture the main features of our simple example.

When the nature of the system under consideration is less clear, however, learning about it consists of more than updating some prior probability distribution in a given probability space that describes all the possible system states. In experimental, explorative research, especially when one tries to understand a new phenomenon, no adequate probability space may be known. In scientific discovery, the most interesting things happen when the phenomena at hand force one to extend, or even completely revise, the framework that one may have used in an initial attempt to describe the system.²⁸ In cases of middle-level difficulty, this can mean that we are not sure which variables (degrees of freedom) are relevant for a certain phenomenon, and that we may need to take into account different or more variables. This assumes, however, that the variables that mathematically characterise the system are already *disclosed*—only waiting to be included in the description, as it were. This assumption may be justified in many practical cases, for example in the context of big-data analysis and the study of correlations. It fails, however, in the context of genuine explorative research, e.g., when a new concept needs to be introduced in order to explain a given phenomenon.

In such explorative contexts, which exemplify our third level of difficulty, a formal notion of learning is hard to establish. As an informal statement, the Bayesian characterisation of learning as “updating one’s prior under newly acquired information” is still a good starting point. However, “one’s prior” then cannot be “one’s prior probability distribution”—there may be no such thing, it may be only

²⁶ The literature in this field is considerable. See, e.g., Paris and Řeháček (2004), Gühne and Tóth (2009), and Friis et al. (2019).

²⁷ A recent tomography experiment that employs the QBist resource of SIC-POVMs is reported by Stricker et al. (2022).

²⁸ For an in-depth account of the phenomenology of experimentation with a focus on its explorative aspects, see Rheinberger (2021a,b).

partially established, or it may not be trusted. In our view, a modelling perspective is helpful at this point because it offers a much broader framework within which learning can be understood as an interactive endeavour. A scientific agent updates her prior *model*, reacting to novel information or, more broadly, to novel experiences that in turn are interpreted in the light of that prior model. If the model changes, the information conveyed by one and the same phenomenon may also change: there can be new information before a familiar background, but there can also be a new background behind familiar information. This opens a space for associations and for transfer learning. In terms of Pearl's ladder of causation and our parallel ladder of attribution discussed in Sect. 2.3.2, the scientific agent here truly operates on the third rung, as a model-building agent who can use counterfactual thinking.

We illustrate the three mentioned levels of difficulty of the process of scientific modelling by way of a quantum-mechanical example from theoretical chemistry.

4.4.1 A Case Study of Modelling: Computing Chemical Binding Energies

One of the main problems of theoretical chemistry is to compute the binding energies of molecules. Precise knowledge of the energy eigenvalues of the molecular structure Hamiltonian would allow one to determine many chemical properties of interest. For example, if one knows just the energy of the ground state for different configurations of the atomic nuclei, one can already learn something about the geometry of the molecule (e.g., bond lengths and angles) and its thermodynamic properties in chemical reactions (e.g., reaction enthalpy). It is, however, computationally challenging to solve the complex many-body problem of finding the energies of the ground state and of the excited states for larger molecules, even if one has full knowledge of the Hamiltonian of the molecule that describes the electronic interactions between all of its constituents. As the complex many-body problem occurs in a wide range of applications in chemistry, condensed matter physics, and material science, efficient methods for solving it have great practical import. The required quantum mechanical computations cannot be carried out analytically, so that one has to resort to approximate and often iterative numerical methods.

One common approach is to use variational methods, which can be viewed as fairly simple iterative schemes of modelling and updating. A variational approach typically starts with some ansatz wave function from a heuristically defined basis set. This ansatz is then refined in a series of iterations, leading to better and better estimates of the ground state (that is, upper bounds of its energy). One theoretical basis for these methods is the Raleigh-Ritz variational principle,²⁹ which can be

²⁹ A central tool that is used in this context is the Born-Oppenheimer approximation. See Head-Gordon and Artacho (2008) for a compact survey that also discusses more advanced computational methods.

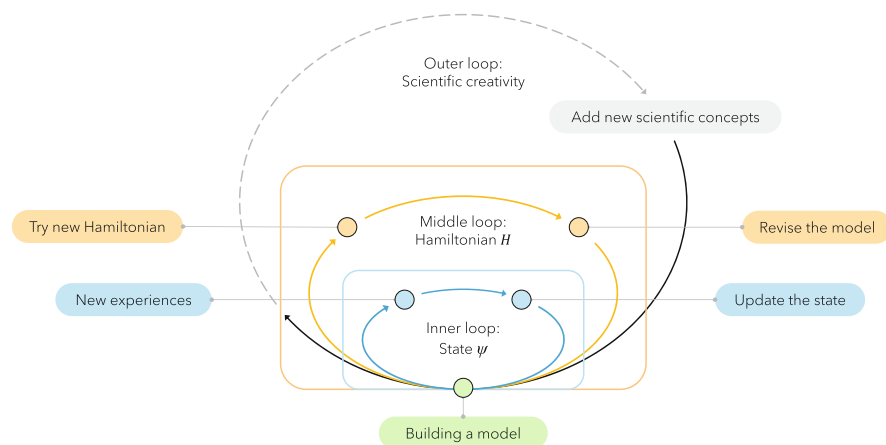


Fig. 4.2 Scheme of learning and modelling in science, showing three modelling loops that correspond to the three levels of difficulty of the modelling task as described in the main text

seen as a concise mathematical framework for recursively “updating one’s prior”, in QBism jargon, within a circumscribed set of candidate wave functions. That updating of the wave function is based entirely on the result of computations and need not involve any measurements on the chemical system itself. Recently, with the advent of quantum computing resources, new methods have been developed, known as variational quantum-classical eigensolvers (McClean et al. 2016), where the computational procedure is interspersed with measurements on a physical system that serves as a quantum simulator of the molecular interactions (O’Malley et al. 2016; Hempel et al. 2018). The results of the measurements provide estimates of the expectation values of certain observables and are consequently used to update the previous approximation of the molecular ground state wave function.

The procedures that we have described so far illustrate the simplest of the three levels of difficulty that we mentioned above. Figure 4.2 provides a schematic view of an agent’s possible ways of learning through the use of QM for all these three levels. In that Figure, variational iteration corresponds just to the inner loop. This process can be fully automatised (Kokail et al. 2019), including the just mentioned interaction of a classical computer with a quantum simulator as part of the variational procedure. Once set up, it needs no human intervention to converge to an estimate of the ground state wave function and its energy.

In general, there is no guarantee that the method of variational iteration will provide the correct ground state wave function. The iteration of this inner loop of Fig. 4.2 may converge, but deliver a wave function whose predictions for certain properties—such as the ground state energy or various binding energies—disagree with observed or inferred properties of the real chemical system. In such a case, one needs to revise the starting assumptions. This could mean to change the set of candidate wave functions over which the variation is carried out, or it could mean to change the Hamiltonian itself. In the latter case, automated iterative estimation procedures are known as Hamiltonian learning schemes.

In Fig. 4.2, Hamiltonian learning corresponds to the second loop, where one revises one's beliefs over a larger set of prior assumptions, including the parameters and possibly also the form of the structure Hamiltonian. This is required, for example, when one knows all constituents of the system, but not the precise strengths of their interactions, which are represented via the coupling parameters in the Hamiltonian. One example of such a problem is finding the spatial structure of large molecules such as proteins. In a protein folding problem, the sequence of amino acids in the protein is precisely known, but the detailed spatial configuration, which depends on the strengths of the Hamiltonian couplings and which determines many of the protein's biochemical properties, is unknown. There have recently been spectacular advances in this area based on sophisticated machine learning techniques (see, e.g., Senior et al. 2020; Jumper et al. 2021), even leading to a share in the 2024 Nobel Prize in Chemistry.

A somewhat different situation arises when the composition of the system or of its environment is not fully known—for example, when it is not known how many environmental degrees of freedom of a certain type a system interacts with. One concrete example of such a setting concerns nitrogen-vacancy (NV) centres in diamond, where an electron spin located at the NV centre interacts with a set of neighbouring nuclear spins whose precise number and interaction strengths may not be known. In a recent study of this problem by Gentile et al. (2021), a learning agent constructs a composite Hamiltonian model out of an overall set of candidate Hamiltonians, using Bayesian inference and parameter estimation. In terms of Fig. 4.2, this procedure involves both the second loop inside the box (by updating the overall structure of the Hamiltonian) and the inner loop (by parameter estimation of its couplings). While the method differs in many respects from the variational approximation of the ground-state wave function discussed above, the overall framework is similar in that the full set of candidate Hamiltonians is fixed and pre-specified.³⁰

Problems at the highest, third level of difficulty do not fit into a fixed iterative scheme, and so we picture them as reaching outside the boxes of Fig. 4.2. In genuine scientific discovery, our lack of understanding of a phenomenon or of a given system can go far beyond the mere lack of knowledge of the precise values of certain parameters or of the number of terms in an otherwise well specified Hamiltonian. The anomalous Zeeman effect, discussed briefly in Sect. 3.1.2, provides a historical example of a fundamental lack of understanding. That phenomenon, the splitting of

³⁰ There is a continuous transition between learning the overall form of the Hamiltonian—i.e., how many and which nuclear spins interact with the NV centre—and learning the parameters describing these interactions—i.e., the strength of the interaction of each nuclear spin in the NV environment. In principle, one could consider a “large enough” Hamiltonian that includes all possible interactions between some maximum number of environmental spins and the NV centre, and then simplify the problem to that of learning the values of the interaction parameters. Many of these would turn out to be vanishing to a very good approximation. Such a straightforward procedure based only on parameter learning and optimisation would involve just the inner loop of Fig. 4.2. It would, however, typically be very inefficient and prone to getting stuck in local minima.

lines in atomic emission spectra into an even number of sub-lines in the presence of a static magnetic field, could only be explained after a new physical property, spin, had been postulated and added to the description of the electron and of the atomic structure. It took over 25 years, from the discovery of the even-numbered line splitting in the late 1890s to the theoretical postulation of spin in the mid-1920s, for this conceptual leap to be made. The postulation of spin as a separate property of the electron—over and above its charge, its mass, and its point-particle nature—was a novel idea, and it introduced a new “quantum number”, a new degree of freedom in the description of an electron’s place in an atom that had not been known or even been considered before.³¹ Different from the examples discussed above, spin was not a candidate degree of freedom out of a predefined set that had been available to be included for the refinement of some Hamiltonian model if needed. Phenomenologically speaking, it had not yet been *disclosed*, and was thus not available as part of the theorists’ toolbox at the time. Spin truly came from somewhere else, from outside of the box.

It is not clear how one can formalise such a discovery process and whether a machine could possibly invent spin to explain the splitting of atomic line spectra. The problem cannot be cast as an optimisation task in a natural way because spin was simply not available as a candidate degree of freedom, neither for the refinement of some model nor for the extension of some variational space. On the other hand, the conceptualisation of spin did not happen in a historical conceptual void, but in a context that provided anchors for analogies—angular momentum was known from mechanics and its mathematics was familiar, and it was known how electromagnetic phenomena (currents in wires, or moving charges generally) can give rise to forces or torques. Bohr’s model of the atom furthermore had familiarised physicists with the assumption of quantisation as a brute fact, and the intensive study of atomic spectra at the time, briefly described in Sect. 3.1.2, provided further relevant concepts, partial models, and numerical formulae. Experiments and models by generations of working scientists, from Ørsted and Ampère to Lorentz and Bohr, Landé and many others thus formed the background before which the notion of spin could be introduced.

This process of learning and innovation, which includes creative modelling and discovery, is an outstanding example of collective human scientific agency. It is not yet clear how, and to what extent, such a process could be realised by machines exhibiting artificial agency. That question is a topic of active research in AI aiming at the automation of science. In Chap. 5 we describe some first steps for understanding such a process within our framework of Projective Simulation.

³¹ As mentioned in Sect. 3.1.2, the idea of introducing such a degree of freedom was even actively suppressed, e.g., by Pauli. It is perhaps no coincidence that the paper by Uhlenbeck and Goudsmit (1925) that announced the idea was written by two graduate students who, according to their supervisor Ehrenfest, had yet no reputation to lose (Goudsmit 1971).

4.4.2 *Agents as Modellers, Models as Tools*

Agents can enter the modelling loop depicted in Fig. 4.2 from different starting points when they try to understand a system: from the experience of a puzzling phenomenon in the observation of nature or in experimenting, driven by a theoretical problem arising in an already-developed theory, via analogies to other models with which they are already familiar, or simply in order to try something out. Depending on the case, the modelling loop can be iterated almost without empirical input, as when one takes a molecular structure Hamiltonian at face value and solves the resulting quantum many-body problem numerically using variational methods, or in constant interaction with a data stream from the system of interest, as in adaptive control.³² Parts of the modelling loop can in many cases be automatised to a high degree. And just as there are different ways of entering the modelling loop, there are also various ways of exiting it: the model may be good enough for the purpose at hand, or the best that can be had given time or resource constraints, or the agent may simply have other things to do and leave the model as it is.

The success of a modelling endeavour depends on the context in which the agent engages in modelling, and on her aims. A successful model for the ground state energy of a molecular substance alone will not help much to understand the substance's toxicological profile, let alone its market value. To be sure, the substance also has such properties, and they can also be modelled, but in different contexts and with different methods. A model is good when it is fit for purpose, and that can mean very different things even with respect to one and the same target system. When a scientific agent familiarises herself with a system in order to gain understanding or control, the modelling process is, first and foremost, the creation of a tool. By modelling the system, the agent gets a handle on it, often in a fairly literal sense: she can learn to manipulate the system in various ways, to predict what will happen as a result of certain interventions, or to get the system to show some desired effect. As a signal that sufficient understanding has been gained, the agent will often stop experimenting on the system—even though the range of possible experiments is unlimited, and new experiences could always be elicited. At some point, the tool is fit for purpose, good enough, or simply the best that can be had given external constraints. That tool can then be put to use in various ways.

In philosophy of science, models are mainly characterised in terms of *representations* of the target system (Frigg & Hartmann 2025). While this approach is often useful, characterising models as *tools* is more adequate from our agency perspective.³³ Tools mediate between an agent and the world, pointing both ways;

³² Adaptive control will play an important role in the stabilisation of complex many-body systems such as a quantum computer. In Sect. 5.4.1 we discuss how adaptive control, leveraged by learning as offered by Projective Simulation, can be used to optimise quantum error correcting codes tailored to specific hardware implementations.

³³ See also Knuuttila and Merz (2009), who affirm the “tool-like character” (p. 150) of models based on their pragmatic approach to modelling.

this corresponds well with the slogan of “models as mediators” that also has currency in philosophy of science.³⁴ And tools, like models, can be more or less fit for a purpose: just like the Hamiltonian model of a molecular substance has got nothing to offer for estimating its market value, a tool that is ideal for one purpose (like driving in a nail) can be ridiculously bad for another purpose (like mending a sock).

From a representational point of view, it is often pointed out that models incorporate idealisations or distortions, which accounts for their fitness or unfitness for certain uses. We agree that this view is often useful, but it betrays an ideal of undistorted representation that should not be accepted uncritically. Taken to its extreme, the idea of an undistorted, completely accurate representation of a target system suggests that an ideal model should be an exact copy, or a map at the scale of 1:1. Borges (1946) succinctly exposes the general futility of this idea. There are indeed a few select circumstances in which having an exact copy of a system is useful: exact copies of the materials that the astronauts had available on board famously helped ground control avert the near-catastrophe of Apollo 13, and NASA still keeps exact copies, e.g., of its Mars rovers for testing interventions on them. But in most cases, the target system is a poor tool for understanding itself. If a system is not yet understood, this means that one does not yet have a handle on it, and the purpose of modelling is exactly to provide such a handle. Rather than viewing this process as the creation of a distorted copy of the system, we believe that it is more fruitful to view this process as the interactive and adaptive creation of a tool that allows an agent to deal with the system—of course, always with its purpose in view.

Models as tools are created in a learning context; they are both products and vehicles of learning, iteratively refined through the resistance of the world. We already mentioned that models thereby provide a perspective on the phenomena that the target system exhibits. This means that the iterative modelling loop of Fig. 4.2 can integrate already familiar empirical data in a new light if the model is updated, adding to the flexibility of the modelling framework. The friction that provides normative guidance for a next iteration of the modelling loop can also come from different sources—empirical, theoretical, or practical. And when a model is put to use, this does not mean that modelling has stopped once and for all, or that nothing more could be learned, or that the true ontology of the target system has been found out. It is not even necessary to assume that the notion of a target system’s true nature or ontology makes good sense. Models are always perspectival. As such, they are answerable to standards of empirical adequacy, aiming to capture aspects that the system really has. But this does not mean that there must be one master-model that incorporates all perspectives and which represents the system as if via a duplicate.

³⁴ In the editorial introduction to their classic collection of papers, Morgan and Morrison (1999, 10, emphasis in original) describe models as “mediating instruments” and even as “*autonomous agents*” in modern science. There are many parallels to what we have to say here, but the focus of *Models as Mediators* is different from ours. Morgan and Morrison view models as mediating between theory and world (or theory and data), not between a scientific agent and the world, and the notion of agency assumed in their characterisation of models as agents is very thin.

Modelling has to integrate the real resistance of the world, but it is not committed to a naïve realism about the target system. Updates of models are always possible, and scientific agents are generally free to choose to enter another round of the modelling loop, or to drop a model as unfit for purpose and to start anew.

4.4.3 *Modelling and QBism*

In our view, a modelling perspective provides a helpful option for QBism to broaden its focus. QBism's current main slogan is that “quantum mechanics is a tool any agent can use to evaluate his probabilistic expectations for his personal experience” (Fuchs and Schack 2015; see also Fuchs et al. 2014). In our critique of what we perceive as unnecessary limitations of this view, we pointed out that agents need not have any probabilistic expectations to begin with, and that the reliance on incorrigible personal experiences is epistemologically unsustainable (Sect. 4.2.2). A modelling perspective is helpful here because even those—admittedly important—uses of QM of which the QBist slogan is true are also instances of modelling, while—as we just argued—other scientific uses of QM to which the slogan does not apply can still be brought under the modelling umbrella. Furthermore, we believe that a modelling perspective is helpful for developing the idea of participatory realism with a view to a structured understanding of both the participants and the reality they interact with, thus filling another lacuna of current QBism.

Modelling and QBist Updating The QBist slogan is a fitting description of the inner modelling loop depicted in Fig. 4.2. In state tomography, for example, each successive measurement on (i.i.d. copies of) the system can be used to update one's model of the system's quantum state, which in turn anchors further probabilistic expectations. The agent's experiences here correspond naturally to perceived measurement outcomes.³⁵ The modelling process ends, for example, when a context-dependent convergence criterion is reached, or when the agent runs out of time or resources, so that no further experiences of the system are forthcoming. To give a maximally simple example, when an agent uses a z -oriented Stern-Gerlach device to measure the spin projection of silver atoms from a given (i.i.d.) source and has found 20 consecutive atoms in the upper channel and none in the lower channel, she can be fairly confident that the source produces atoms in the spin state $|\uparrow, z\rangle$. If she starts out from the “no information” assumption that the

³⁵ As we argued in Sect. 3.3.2.2, a measurement outcome presupposes irreversibility. At this point we disagree with the QBist approach, which tries to steer clear from all ontological pronouncements. For our present context, however, this difference is immaterial. We agree that for the simple example in question, the modelling loop can be fully described in terms of an agent's experiences.

source is unpolarised, i.e., that the atoms should be described by the mixed state

$$\rho = \frac{1}{2} (|\uparrow, z\rangle\langle\uparrow, z| + |\downarrow, z\rangle\langle\downarrow, z|), \quad (4.4)$$

then each successive registration of an “up” outcome will shift her state estimate in accordance with the Bayesian update rule.

To see how this works, assume that the agent is confident that the source delivers atoms prepared identically and independently, each of them in one of the two pure states $|\uparrow, z\rangle$ or $|\downarrow, z\rangle$, and that the Stern-Gerlach device really measures in the z -direction. This confidence must build on various assumptions, for example, that the source emits atoms with their spin polarised along only one axis (for example, because the source is itself a Stern-Gerlach device), that the source and the Stern-Gerlach device used for the measurement are stably mounted and precisely aligned, that there are no additional magnetic fields in the region near the source, and so on. The probabilities $\Pr(\uparrow, z) = \beta$ and $\Pr(\downarrow, z) = 1 - \beta$ with which the atoms leave the source are typically assumed to be fixed (i.e., time-independent) but may initially be unknown to the agent. If the value of β were known, the agent would describe the atoms that leave the source by the mixed state $\rho(\beta) = \beta|\uparrow, z\rangle\langle\uparrow, z| + (1 - \beta)|\downarrow, z\rangle\langle\downarrow, z|$.

A Bayesian approach to estimating the value of β is similar to estimating the bias of a coin in a classical coin-flipping experiment. Without any experimental data, the agent’s prior probability distribution for the bias β will be $w^{(0)}(\beta) = 1$.³⁶ Correspondingly, the atoms are described by the mixed state

$$\rho^{(0)} = \int_0^1 d\beta w^{(0)}(\beta) \rho(\beta) = \frac{1}{2} |\uparrow, z\rangle\langle\uparrow, z| + \frac{1}{2} |\downarrow, z\rangle\langle\downarrow, z|, \quad (4.5)$$

which agrees with Eq. (4.4). Suppose now that the source emits every atom such that it will end up in the upper channel of the agent’s Stern-Gerlach device (for example, because the source itself consists of the upper channel of a z -oriented Stern-Gerlach device) but the agent does not know this. In the jargon of coin flipping, this would mean that the source is *maximally biased*, like a coin that always shows heads. How does the agent adapt her description over time? After detecting the first atom in the upper channel, the agent can use Bayes’s formula,

$$\Pr(\beta|\uparrow, z)d\beta = \frac{\Pr(\uparrow, z|\beta)w(\beta)d\beta}{\Pr(\uparrow, z)}, \quad (4.6)$$

to improve her estimate of the value of β , i.e., to update the prior probability density $w(\beta) = w^{(0)}(\beta)$ with the posterior probability density $\Pr(\beta|\uparrow, z)$. Together with

³⁶ Note that $w^{(0)}(\beta)$ is a *probability density*. That is, the probability for finding the bias of the source to be somewhere in the (infinitesimal) interval $[\beta, \beta + d\beta]$ is equal to $w^{(0)}(\beta)d\beta = d\beta$.

her knowledge of quantum mechanics, in the form of the Born rule (Eq. 3.45),

$$\Pr(\uparrow, z|\beta) = \text{tr}(\rho(\beta)|\uparrow, z\rangle\langle\uparrow, z|) = \beta, \quad (4.7)$$

this results in a new estimate of the value of β in the form of the updated probability density $w^{(1)}(\beta)$, such that $w^{(1)}(\beta)d\beta = \Pr(\beta|\uparrow, z)d\beta = 2\beta d\beta$. Iterating the procedure, it is easy to see that, after a sequence of n successive detections of atoms in the upper channel, the agent will obtain $w^{(n)}(\beta) = (n+1)\beta^n$. Accordingly, the updated mixed state for the atoms leaving the source is given by

$$\rho^{(n)} = \int_0^1 d\beta w^{(n)}(\beta)\rho(\beta) = \frac{n+1}{n+2}|\uparrow, z\rangle\langle\uparrow, z| + \frac{1}{n+2}|\downarrow, z\rangle\langle\downarrow, z|. \quad (4.8)$$

Having found 20 consecutive atoms in the upper channel, as in our introductory story, the agent will thus assign a probability close to 95% to finding the next atom again in the upper channel.

The same argument can be repeated with the somewhat weaker assumption that the source emits the atoms with their spin pointing in a random direction, that is, with each atom prepared independently in some state $|\uparrow, \vec{n}_\Omega\rangle$ where the unity vector $\vec{n} = \vec{n}_\Omega$ points into a random direction specified by the spherical coordinates $\Omega = (\vartheta, \varphi)$. In this case the agent would describe the atoms leaving the source by the initial state

$$\rho^{(0)} = \int d\Omega w^{(0)}(\Omega)|\uparrow, \vec{n}_\Omega\rangle\langle\uparrow, \vec{n}_\Omega| = \frac{1}{2}(|\uparrow, z\rangle\langle\uparrow, z| + |\downarrow, z\rangle\langle\downarrow, z|), \quad (4.9)$$

with the prior (flat) probability density $w^{(0)}(\Omega) = \frac{1}{4\pi}$ for the spin direction $\vec{n}_\Omega$. As one can easily see, the resulting description of the atoms is the same as in Eq. (4.5). The proof of the second equality in (4.9) is readily obtained from Eq. (3.90). Successive Bayesian updating of the probability density $w^{(n)}(\Omega)$ (which here describes the emission profile of the source as estimated in step n), for a sequence of n successive detections of atoms in the upper channel, leads to a distribution that is peaked at $\vartheta = 0$, i.e., in the z -direction.

Such a simple example of learning can still anchor a variety of agential behaviours. If the agent in our example all of a sudden starts registering only “down” outcomes, she may be surprised, and she may discard her previously well-confirmed assumption that the source was stationary, leading her, perhaps, to start from scratch with the mixed state of Eq. (4.4). Holding on to the simplest Bayesian recipe will not be a particularly good algorithm here, as it would take a long time for the agent to arrive at a stable pronouncement that the source now produces atoms in the spin state $|\downarrow, z\rangle$. More refined estimation methods can lead to better convergence in such circumstances (see, e.g., Granade et al. 2016). Looking ahead at our modelling

in Chap. 5, one could even include a rule of forgetting. QBism is by no means committed to the most simplistic version of updating.³⁷

When modelling involves not just state estimation but also Hamiltonian learning, as in the examples illustrating the middle level of difficulty in Sect. 4.4.1, the QBist slogan becomes stretched, amounting to the idea that the middle modelling loop of Fig. 4.2 is in fact just an instance of the inner loop. It may indeed be possible to view the modelling efforts in such cases as updates in a large enough probability space, but quite often such a perspective will be available only after the fact, when some creative choices regarding the necessary variables or the size of the parameter space have been made. Furthermore, as we pointed out (see note 30), optimisation within such a large space is often an inefficient and impractical method. As a description of what goes on when a scientific agent engages in the middle level modelling loop, the QBist approach of updating probabilistic expectations appears too limited.

This is all the more so when actual explorative research is concerned, which may invoke all three loops depicted in Fig. 4.2. As we argued in Sect. 4.4.2, modelling is generally best viewed as the creation and iterative refinement of tools by the agent, in interaction both with the world, which is thereby disclosed for the agent, and with other agents facing similar challenges. While we thus stress the social dimension of the scientific endeavour, we agree with QBism that a personal perspective may also be helpful and in fact may be necessary. The models created as tools may well be quite personal. Scientists have their individual quirks and preferences when it comes to notation, terminology, preferred analogies, or order in experimentation. In this respect, models are no different from many other tools we use. Glasses, for example, are individual, custom-fit tools for getting around in the world. If you are

³⁷ The appropriate generalisation of the simple Bayesian method described above will depend on the situation under consideration. For example, there could be uncertainty regarding the alignment of the source. (Consider again, for simplicity, the binary case where the source emits the atoms only in two possible states $|\uparrow, z\rangle$ or $|\downarrow, z\rangle$.) It could be that it is not precisely aligned in the z -direction but with a little offset, for example, into the zx -plane, such that the atoms that leave the source are prepared in an eigenstate $|\uparrow, \vec{n}_\alpha\rangle$ of the spin pointing in the direction $\vec{n}_\alpha = \cos(\alpha)\vec{e}_z + \sin(\alpha)\vec{e}_x$, i.e., in an eigenstate of the observable $\vec{n}_\alpha \cdot \hat{S} = \cos(\alpha)\hat{S}_z + \sin(\alpha)\hat{S}_x$. The same could also happen if the source is correctly aligned in the z -direction, but there are additional magnetic fields acting on the atoms after they have left the source. These could be unknown stray fields, representing experimental imperfections, or additional control fields that are applied on purpose. We described an example of such a scenario in Sect. 3.3.3.3 and we will revisit a similar one in Sect. 5.4.3 when discussing the use of PS learning agents in quantum experiments. In our present example, the prior for the offset angle α would be represented by some distribution $w^{(0)}(\alpha)$ over the interval $\alpha \in [0, 2\pi)$. The corresponding mixed state for the atoms that leave the source is then given by $\rho^{(0)} = \int_0^{2\pi} d\alpha w^{(0)}(\alpha) |\uparrow, \vec{n}_\alpha\rangle \langle\uparrow, \vec{n}_\alpha|$, which for the extreme case of a flat prior, $w^{(0)}(\alpha) = 1/(2\pi)$, would correspond to Eq. (4.4). The efficiency with which the agent learns (or improves its estimate of) the offset will depend, for example, on the type and number of measurements it performs on the outgoing beam of atoms. If we consider again the simplest case where the offset is zero (correct alignment and no stray or other additional fields) and the agent measures the spin only in the z -direction, one can show that the estimate of the offset develops into a distribution that peaks at $\alpha = 0$. In Sect. 5.4.3, we will revisit a similar problem as an application of Projective Simulation with a focus on adaptive learning.

short-sighted, you will not gain much from borrowing your long-sighted friend's glasses for a bicycle tour. You may, however, profit from these same glasses when you borrow them to light a camp fire. Similarly, a scientific agent using QM to develop a model for a system she is familiarising herself with may end up with a model that is quite idiosyncratic, but which helps her devise novel experiments.³⁸ When she wants to communicate that model in a publication, she will need to make an effort to rework it to fit into the existing scientific discourse, so that others can familiarise themselves with the system as well. This process of reworking models in a social context is yet another aspect of creative scientific agency. It may well be that in this process of redescription, the modelling effort is ultimately pictured as an instance of just the inner modelling loop of Fig. 4.2. After all the work is done, a clean description may be available in which the necessary concepts for describing the system in question are all present, the system's Hamiltonian has been worked out, and the only thing left to do is to estimate some interaction parameters. But such a publication- or textbook-style presentation should not distract from the fact that the real scientific work has a much more open texture.³⁹

Participatory Realism We believe that our modelling perspective can help to develop the notion of participatory realism, which in current QBism mostly appears as a programmatic vision far removed from the QBist's technically demanding research into the scope and limits of probabilistic updating. Modelling can bridge this gap. In building models, agents interact with and participate in worldly happenings. We stressed that modelling does not commit one to an ontology, much less to naïve realism about the system one interacts with: models are perspectival and are built for a specific purpose, while reality affords many different and incompatible purposes and perspectives. Building a model is an iterative process of learning in which the world exerts resistance, providing normative guidance for updating the model. Such resistance may well be in the form of an experience either conforming with or going against an agent's probabilistic expectations, as assumed in current QBism. But there are other forms of resistance as well, and a modelling perspective provides the space within which different reactions to such resistance, corresponding to different types of learning, can be admitted. Modelling agents participate in worldly happenings in many different ways.

As to realism, we affirm that modelling is a world-disclosing and therefore realistic endeavour—not in the sense of providing the one true view of the world, but in the sense of outfitting different agents with different, yet normatively constrained perspectives on the one world we live in. As we will argue in the following, from our modelling perspective it is also natural to take agency itself into view, a worldly phenomenon with which we are intimately familiar: Scientific agency as modelling can be extended to model scientific agency itself.

³⁸ For a good example of such idiosyncratic modelling, consider Kamerlingh Onnes's efforts to understand the entirely novel phenomenon of superconductivity discovered in his lab in 1911. See Van Delft and Kes (2010) for an overview and for references to original sources.

³⁹ See Hacking (1983) and Rheinberger (2011b).

4.5 Modelling Scientific Agency

In contradistinction to natural phenomena such as superconductivity, there is no need to *familiarise* ourselves with the phenomenon of agency through modelling—agency is part and parcel of our everyday dealings in the world, and even a precondition of experiencing any other worldly phenomenon. This does not mean, however, that agency modelling would be futile. Far from it. In discussing agency in Chap. 2, we already stressed the importance of a deeper understanding of agency as a real phenomenon in the world, and in our description of scientific agency in this Chapter, we provided some ingredients for such an understanding. In the following Chap. 5, we will discuss at length a concrete agency model, Projective Simulation, which incorporates memory, learning, and decision finding in a modelling framework. That model is meant both as a theoretical contribution towards a deeper understanding of (scientific) agency, and as a concrete practical proposal for how artificial agents could be realised.

Artificial agency is currently a vision, not a reality.⁴⁰ But there are, at the current state of technological development, many ways in which the computerisation and automation of science is advancing. Computers have become versatile and in fact indispensable tools for all types of scientific activity, from literature search and the writing of texts via data storage and data analysis to the full automation of fairly complex tasks, both theoretical and practical. We mentioned the full automation of quantum-classical eigensolvers for estimating molecular ground state energies in Sect. 4.4.1; the even broader vision of fully automatic “self-driving labs” is generating research proposals and concrete business models, e.g., in the life sciences and in chemistry (Häse et al. 2019; Rapp et al. 2024). Machine learning methods drive further advances of automation in many other areas of science, from data analysis and experimental control to experimental design and tools for enhancing the creativity of human brain-storming. So far there are no artificial systems that would count as artificial agents doing science. But efforts towards building them are picking up momentum. In our view, the development of artificial agency must be based on physically realistic and philosophically well-grounded modelling. Such modelling will also contribute to the development of QBism’s vision of participatory realism by giving internal structure to the central QBist notion of an agent as a user of scientific theories such as QM.

4.5.1 Modelling Agency, Modelling Modelling

Since we view scientific agency as centrally a modelling endeavour, to model scientific agency means to model the scientific agents’ modelling. Our description

⁴⁰ See Van Lier (2025) for a recent methodological discussion of agency modelling that starts from this premise.

of scientific modelling as an iterative, interactive and world-disclosing activity in Sect. 4.4 therefore already provides some of the main components for modelling scientific agency. Further important aspects of agency modelling come from our explication of agency that we spelled out in Sect. 2.6.

As we argued at length in Chap. 2, agency is an ontologically emergent phenomenon in our world. As such, it presupposes the basic indeterminism of the matter that agents consist of. QM is a promising framework in this respect because of the fundamental indeterminism of the basic dynamics of natural processes that QM describes (Sect. 3.6). An agency model can make use of this aspect of QM in various ways, in line with the various ways of integrating agency and QM discussed above (Sect. 4.1). As we argued there, rather than trying to model agents as quantum systems themselves, it seems best to take some core features of quantum dynamics—such as indeterminism with adjustable probabilities—as building blocks for an agency model whose overall structure is geared towards the other aspects of agency, and especially of scientific agency, that we highlighted in our list of features in Sect. 2.6.1. These aspects include embodiment, settling, flexibility, and adaptivity, which are underwritten by the use of QM as a modelling framework. Additionally, they include wanting, learning, and historicity, which do not have corresponding counterparts in the theory of QM. We hold, however, that they can be integrated into an agency model via structural analogies. Importing structure from QM in this higher-level way also does justice to the insight, argued for in Sect. 2.6.1.ii, that agency presupposes indeterminism on at least two different levels: at the level of the physical matter, described by QM, and at the level of the body, where that matter is organised in such a way that further possibilities for organisation are still left open for the agent to settle. Correspondingly, core features of quantum dynamics are thus useful in agency modelling both at the level of the underlying matter and at the level of more abstract aspects such as historicity, learning, or the agent’s memory structure.

These aspects are important for agency quite generally, as stressed in Sect. 2.6. For scientific agency, however, we can be more specific: much of the learning and adaptation that is relevant here can be described as modelling, and a scientific agent’s interests, which influence what she wants, are also oriented towards modelling and making a model fit for its purpose. As we argued in Sect. 4.4, scientific agency to a large extent aims at building models as tools for understanding nature and for getting around in the world, so that a scientific agent’s wanting is naturally focused on predictive success and precise control.

4.5.2 *Towards Projective Simulation*

What we have laid out in this Chapter agrees with the QBist doctrine in a number of respects. The most important one, in our view, is the focus on learning from novel experiences as the central means for an agent to get around in and adapt to its environment, be it for scientific or for other purposes. This means that an agent’s

learning over time must take into account the feedback that the agent receives from the world. Such feedback may simply consist of a prediction error signal, which in the context of the inner modelling loop of Fig. 4.2 may then be used for parameter optimisation—for example, via gradient descent. Scientific learning is, however, not limited to such simple updating in closed and formally specified settings, and here our own approach to learning goes beyond QBism.

In Sect. 4.4.3 we already argued that we need to broaden QBism's view on learning because scientific agency as modelling reaches beyond the simple learning scenario of updating in a given probability space. An agent may not have a prior probability distribution or a fully worked out set of environmental parameters that it could update, but rather simply try to get better at navigating its environment in a sensible way. We describe this as the agent's developing and refining its inner *model*, its tool for getting along in the world. In the following Chap. 5 we describe a versatile memory structure for an agent that allows for the update of probabilistic expectations, but also for a broader class of updates via the generation of new options for action and new conceptual structures. Such an agent's inner model may have a structure that explicitly mimics or represents features of the environment, but this is by no means necessary. If an agent gets along, we can ascribe to it a model at least in the minimal sense of a tool for navigating its environment: A successful agent, simply by virtue of being successful, proves that it has developed a useful tool for dealing with the world.

That idea is well captured by the agency model of Projective Simulation that we already mentioned in Sect. 2.7.1 and which we will discuss in detail in the following two Chaps. 5 and 6. Having available an inner model as a vehicle for learning means that an agent need not try out every option it has, nor stumble through its available actions. The agent can, rather, *simulate* its actions before putting them to the test of reality. The notion of simulation, like the notion of a model, here has a minimal, non-representational sense: The agent need not calculate possible outcomes of its various possible actions, but can use its model to *project* itself into various possible futures.

References

- Adam, C., & Tannery, P. (1904). *Oeuvres de Descartes*. J. Vrin.
- Appleby, M., Flammia, S., McConnell, G., & Yard, J. (2017). SICs and algebraic number theory. *Foundations of Physics*, 47(8), 1042–1059.
- Bäckström, S., & Gustafsson, M. (2017). Skill, drill, and intelligent performance: Ryle and intellectualism. *Journal for the History of Analytical Philosophy*, 5(5), 40–55.
- Bengtsson, I. (2020). SICs: Some explanations. *Foundations of Physics*, 50(12), 1794–1808.
- Bohm, D., & Hiley, B. J. (1993). *The undivided universe: An ontological interpretation of quantum theory*. Routledge.
- Bohr, N. (1934). *Atomic theory and the description of nature*. Cambridge University Press.
- Borges, J. L. (1946). Del rigor en la ciencia. In *El Hacedor*, 1961. Translated by Andrew Hurley, Borges, *Collected Fictions*, Penguin 1998, as “On Exactitude in Science”. Newer translated by Diego Doval as “On rigor in science”. <https://medium.com>, 2017.

- Candlish, S., & Wrisley, G. (2019). Private language. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition. <https://plato.stanford.edu/archives/fall2019/entries/private-language/>
- Cottingham, J., Stoothoff, R., & Murdoch, D. (1984). *The philosophical writings of Descartes*. Cambridge University Press.
- De Finetti, B. (2017). *Theory of probability. A critical introductory treatment* (Combined edition.). Wiley Two volumes in one book. Translated from the 1970 Italian original, *Teoria Delle Probabilita*, by Antonio Machí and Adrian Smith.
- DeBroda, J. B., Fuchs, C. A., Pienaar, J. L., & Stacey, B. C. (2021). Born's rule as a quantum extension of Bayesian coherence. *Physical Review A*, 104:022207.
- DeBroda, J. B., Fuchs, C. A., & Schack, R. (2020). Respecting one's fellow: QBism's analysis of Wigner's friend. *Foundations of Physics*, 50, 1859–1874.
- DeBroda, J. B., Fuchs, C. A., & Schack, R. (2024). QBism's account of quantum dynamics and decoherence. *Physical Review A*, 110, 052205.
- DeBroda, J. B., & Love, P. J. (2022). Quantum and classical Bayesian agents. *Quantum*, 6, 713.
- DeBroda, J. B., & Stacey, B. C. (2018). FAQBism. arXiv 1810.13401.
- Dieks, D. (2022). Perspectival quantum realism. *Foundations of Physics*, 52, 95.
- Dunjko, V., & Briegel, H. J. (2018). Machine learning & artificial intelligence in the quantum domain: A review of recent progress. *Reports on Progress in Physics*, 81, 074001.
- Dunjko, V., Taylor, J. M., & Briegel, H. J. (2016). Quantum-enhanced machine learning. *Physical Review Letters*, 117, 130501.
- Earman, J. (2019). Quantum Bayesianism assessed. *The Monist*, 102(4):403–423.
- Elliott, T. J., Gu, M., Garner, A. J. P., & Thompson, J. (2022). Quantum adaptive agents with efficient long-term memories. *Physical Review X*, 12, 011007.
- Englert, B.-G. (2013). On quantum theory. *European Physical Journal D*, 67, 238.
- Felappi, G. (2023). Propositionalism and questions that do not have correct answers. *Erkenntnis*, 88, 2063–2081.
- Flamini, F., Krumm, M., Fiderer, L. J., Müller, T., & Briegel, H. J. (2024). Towards interpretable quantum machine learning via single-photon quantum walks. *Quantum Science and Technology*, 9, 045011.
- Frauchiger, D., & Renner, R. (2018). Quantum theory cannot consistently describe the use of itself. *Nature Communications*, 9, 3711.
- French, S. (2023a). *A phenomenological approach to quantum mechanics. Cutting the chain of correlations*. Oxford University Press.
- French, S. (2023b). Putting some flesh on the participant in participatory realism. In P. Berghofer & H. A. Wilsche (Eds.), *Phenomenology and QBism: New approaches to quantum mechanics* (pp. 268–289). Routledge.
- Frigg, R., & Hartmann, S. (2025). Models in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2025 edition. <https://plato.stanford.edu/archives/sum2025/entries/models-science/>
- Friis, N., Vitagliano, G., Malik, M., & Huber, M. (2019). Entanglement certification from theory to experiment. *Nature Reviews Physics*, 1(1), 72–87.
- Fuchs, C. A. (2010). QBism, the perimeter of quantum Bayesianism. arXiv 1003.5209.
- Fuchs, C. A. (2017a). Notwithstanding Bohr, the reasons for QBism. *Mind and Matter*, 15(2), 245–300.
- Fuchs, C. A. (2017b). On participatory realism. In I. T. Durham & D. Rickles (Eds.), *Information and interaction: Eddington, Wheeler, and the limits of knowledge* (pp. 113–134). Springer.
- Fuchs, C. A., Hoang, M. C., & Stacey, B. C. (2017). The SIC question: History and state of play. *Axioms*, 6(3), 21.
- Fuchs, C. A., Mermin, N. D., & Schack, R. (2014). An introduction to QBism with an application to the locality of quantum mechanics. *American Journal of Physics*, 82(8), 749–754.
- Fuchs, C. A., Olshanii, M., & Weiss, M. B. (2022). Quantum mechanics? It's all fun and games until someone loses an *i*. arXiv 2206.15343.

- Fuchs, C. A., & Schack, R. (2013). Quantum-Bayesian coherence. *Reviews of Modern Physics*, 85, 1693–1715.
- Fuchs, C. A., & Schack, R. (2015). QBism and the Greeks: Why a quantum state does not represent an element of physical reality. *Physica Scripta*, 90(1), 015104.
- Gentile, A. A., Flynn, B., Knauer, S., Wiebe, N., Paesani, S., Granade, C. E., Rarity, J. G., Santagati, R., & Laing, A. (2021). Learning models of quantum systems from experiments. *Nature Physics*, 17(7), 837–843.
- Gühne, O., & Tóth, G. (2009). Entanglement detection. *Physics Reports*, 474(1), 1–75.
- Glauber, R. J. (1986). Amplifiers, attenuators, and Schrödinger's cat. *Annals of the New York Academy of Sciences*, 480(1), 336–372.
- Glick, D. (2021). QBism and the limits of scientific realism. *European Journal for Philosophy of Science*, 11, 53.
- Goudsmit, S. A. (1971). De ontdekking van de electronenrotatie. *Nederlands Tijdschrift voor Natuurkunde*, 37(16), 386–392. English translation as “The discovery of the electron spin” (by J.H. van der Waals) in *Foundations of Modern EPR*, ed. G.R. Eaton, S.S. Eaton, and K.M. Salikhov, World Scientific, Singapore, 1998, pp. 1–12.
- Granade, C., Combes, J., & Cory, D. G. (2016). Practical Bayesian tomography. *New Journal of Physics*, 18(3), 033024.
- Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S., & Eisert, J. (2010). Quantum state tomography via compressed sensing. *Physical Review Letters*, 105, 150401.
- Hacking, I. (1983). *Representing and intervening*. Cambridge University Press.
- Head-Gordon, M., & Artacho, E. (2008). Chemistry on the computer. *Physics Today*, 61(4), 58–63.
- Healey, R. (2023). Quantum-Bayesian and pragmatist views of quantum theory. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, winter 2023 edition. <https://plato.stanford.edu/archives/win2023/entries/quantum-bayesian/>
- Hempel, C., Maier, C., Romero, J., McClean, J., Monz, T., Shen, H., Jurcevic, P., Lanyon, B. P., Love, P., Babbush, R., Aspuru-Guzik, A., Blatt, R., & Roos, C. F. (2018). Quantum chemistry calculations on a trapped-ion quantum simulator. *Physical Review X*, 8, 031022.
- Hájek, A. (2008). Arguments for, or against, probabilism? *British Journal for the Philosophy of Science*, 59(4), 793–819.
- Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1, 282–291.
- Jerbi, S., Fiderer, L. J., Poulsen Nautrup, H., Kübler, J. M., Briegel, H. J., & Dunjko, V. (2023). Quantum machine learning beyond kernel methods. *Nature Communications*, 14, 517.
- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Žídek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.
- Kane, R. (1998). *The significance of free will*. Oxford University Press.
- Kane, R. (2024). *The complex tapestry of free will: A philosophical odyssey*. Oxford University Press.
- Kewming, M. J., Shrapnel, S., & Milburn, G. J. (2021). Designing a physical quantum agent. *Physical Review A*, 103, 032411.
- Knuuttila, T., & Merz, M. (2009). Understanding by modeling. In H. W. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding. Philosophical perspectives*. University of Pittsburgh Press.
- Kokail, C., Maier, C., van Bijnen, R., Brydges, T., Joshi, M. K., Jurcevic, P., Muschik, C. A., Silvi, P., Blatt, R., Roos, C. F., & Zoller, P. (2019). Self-verifying variational quantum simulation of lattice models. *Nature*, 569(7756), 355–360.
- Ladyman, J. (2023). Structural realism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition. <https://plato.stanford.edu/archives/sum2023/entries/structural-realism/>
- Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48(1), 19–49.

- Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition*, 34, 1667–1675.
- Lukens, J. M., Law, K. J. H., Jasra, A., & Lougovski, P. (2020). A practical and efficient approach for Bayesian quantum state estimation. *New Journal of Physics*, 22(6), 063038.
- McClean, J. R., Romero, J., Babbush, R., & Aspuru-Guzik, A. (2016). The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2), 023023.
- Mermin, N. D. (2017). Why QBism is not the Copenhagen interpretation and what John Bell might have thought of it. In R. Bertlmann & A. Zeilinger (Eds.), *Quantum [Un]speakables II: Half a century of bell's theorem* (pp. 83–93). Springer.
- Millgram, E. (2005). *Ethics done right*. Cambridge University Press.
- Montague, M. (2007). Against propositionalism. *Noûs*, 41(3), 503–518.
- Morgan, M. S., & Morrison, M. (1999). *Models as mediators*. Cambridge University Press.
- O'Malley, P. J. J., Babbush, R., Kivlichan, I. D., Romero, J., McClean, J. R., Barends, R., Kelly, J., Roushan, P., Tranter, A., Ding, N., Campbell, B., Chen, Y., Chen, Z., Chiaro, B., Dunsworth, A., Fowler, A. G., Jeffrey, E., Lucero, E., Megrant, A., Mutus, J. Y., Neeley, M., Neill, C., Quintana, C., Sank, D., Vainsencher, A., Wenner, J., White, T. C., Coveney, P. V., Love, P. J., Neven, H., Aspuru-Guzik, A., & Martinis, J. M. (2016). Scalable quantum simulation of molecular energies. *Physical Review X*, 6, 031007.
- Oreskes, N. (2019). *Why trust science?* Princeton University Press.
- Paparo, G. D., Dunjko, V., Makmal, A., Martin-Delgado, M. A., & Briegel, H. J. (2014). Quantum speedup for active learning agents. *Physical Review X*, 4(3), 031002.
- Paris, M. G. A., & Řeháček, J. (Eds.). (2004). *Quantum state estimation. Lecture notes in physics* (vol. 649). Springer.
- Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.
- Pienaar, J. (2020). Extending the agent in QBism. *Foundations of Physics*, 50, 1894–1920.
- Pienaar, J. (2021). QBism and relational quantum mechanics compared. *Foundations of Physics*, 51(5), 1–18.
- Rapp, J. T., Bremer, B. J., & Romero, P. A. (2024). Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1), 97–107.
- Renes, J. M., Blume-Kohout, R., Scott, A. J., & Caves, C. M. (2004). Symmetric informationally complete quantum measurements. *Journal of Mathematical Physics*, 45, 2171–2180.
- Rheinberger, H.-J. (2021a). On the narrative order of experimentation. In M. Carrier, R. Mertens, & C. Reinhardt (Eds.), *Narratives and comparisons: Adversaries or allies in understanding science?* Bielefeld University Press.
- Rheinberger, H.-J. (2021b). *Spalt und Fuge. Eine Phänomenologie des Experiments*. Suhrkamp.
- Ryle, G. (1949). *The concept of mind*. Hutchinson.
- Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, 23, 807–812, 823–828, 844–849. English translation by J. D. Trimmer, “The Present Situation in Quantum Mechanics”, *Proceedings of the American Philosophical Society* 124(5), 1980, pp. 323–338.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penedones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710.
- Stacey, B. C. (2019). Ideas abandoned en route to QBism. arXiv 1911.07386.
- Stricker, R., Meth, M., Postler, L., Edmunds, C., Ferrie, C., Blatt, R., Schindler, P., Monz, T., Kueng, R., & Ringbauer, M. (2022). Experimental single-setting quantum state tomography. *Physical Review X Quantum*, 3, 040310.
- t'Hooft, G. (2016). *The cellular automaton interpretation of quantum mechanics*. Springer.
- Timpson, C. G. (2008). Quantum Bayesianism: A study. *Studies in History and Philosophy of Modern Physics*, 39(3), 579–609.
- Uebel, T. E. (2007). *Empiricism at the crossroads: The vienna circle's protocol-sentence debate*. Open Court.

- Uhlenbeck, G. E., & Goudsmit, S. (1925). Ersetzung der Hypothese vom unmechanischen Zwang durch eine Forderung bezüglich des inneren Verhaltens jedes einzelnen Elektrons. *Naturwissenschaften*, 47, 953–954.
- Vaidman, L. (2021). Many-worlds interpretation of quantum mechanics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition. <https://plato.stanford.edu/archives/fall2021/entries/qm-manyworlds/>
- Van Delft, D., & Kes, P. (2010). The discovery of superconductivity. *Physics Today*, 63(9), 38–43.
- Van Lier, M. (2025). *Modeling the artificial agent*. Ph.D. Thesis, University of Konstanz.
- Vineberg, S. (2022). Dutch book arguments. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition. <https://plato.stanford.edu/archives/fall2022/entries/dutch-book/>.
- Wheeler, J. A., & Zurek, W. H. (Eds.). (1984). *Quantum theory and measurement*. Princeton University Press.
- Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica*, 43, 99–124.
- Żukowski, M., & Markiewicz, M. (2021). Physics and metaphysics of Wigner's friends: Even performed premeasurements have no results. *Physical Review Letters*, 126, 130402.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 5

Projective Simulation



In this chapter we will give an introduction to the model of Projective Simulation. The main objective of this chapter is to provide an exposition that is both self-contained and serves the other chapters of the book, which focus on agency, quantum mechanics, QBism, and phenomenology, respectively. We choose a level of detail that should be accessible with a basic background in mathematics, comparable to our exposition of quantum mechanics in Chap. 3.

We will first give an informal introduction to the model in Sect. 5.1, by way of describing a hypothetical device that we call a random movie player. We will then represent a more formal account of the model in Sect. 5.2, including the structures that constitute the agent's memory and the rules for deliberation and learning, all of which are illustrated with basic examples. Section 5.3 provides the technical background that is needed to relate PS to standard reinforcement learning, a branch of machine learning. In Sects. 5.4 and 5.5 we discuss some applications of PS agents to problems in quantum information science and beyond. These applications build upon the quantum information background provided in Sects. 3.4.4 and 3.5.4, and they connect to the general discussion of the role of agency in quantum mechanics and in QBism from Chaps. 3 and 4, respectively. In Sect. 5.6, we then present quantum extensions of PS that use quantum resources for certain agential capacities. In Sect. 5.7, we discuss the explainability of the PS agent's decisions and dispositions, relating the PS model—both its classical and quantum versions—to the field of explainable AI. At the end of this Chapter, in Sect. 5.8, we come back to the issue of model-building and agency modelling, which connects to open questions from the previous Chap. 4, and we situate PS as an agency model with reference to the criteria developed in Chap. 2.

5.1 Informal Description of the Model

Imagine a hypothetical device called a *random movie player*. If you upload a movie to the random movie player and then watch that movie again, it will play out a little differently every time you watch it. Typically, there will be small variations of how the characters behave in a given scene and how an individual scene plays out. Sometimes, there may also be larger variations, where entire scenes are deleted or new scenes added, or where the plot of the movie changes.

We envision the random movie player as an interactive device. If you watch a movie repeatedly, the sequence of replays is not completely random. Rather, you can influence the replays by giving feedback to the device. For example, you could rate the replays on some scale to indicate whether you liked them or not. This feedback could be given at the end of a movie, or in shorter intervals while the movie is still playing, i.e., in an online mode.

The random movie player would work like a film editor, operating on the raw footage of a movie, the raw sequence of shots. Each shot itself consists of a number of images or frames. The original movie, as uploaded on the random movie player, thus consists of a selected sequence of shots. However, based on the raw footage and possibly adding shots collected from other, similar recordings or shootings, the player can generate new movies by processing the material in different ways. For example, it could vary the selection, scramble the order, or even modify the content of individual shots.¹ The movies that such a random movie player generates will depend on a number of factors. These include the repertoire of movies (or the corresponding raw footage) that the player can access or has been trained with, the level of randomness with which it combines and alters shots (which could be tuned), and, importantly, on the feedback of the users after or during the roll-outs.²

Projective Simulation (PS), the topic of this chapter, is a model of agency in which the memory of the agent is something like a random movie player. As a model of agency, PS has to integrate the key points from our explication of agency given in Sect. 2.6, i.e., embodiment, indeterminism, wanting, and historicity.

To introduce the PS model, it is useful to start with the standard agent-environment framework of artificial intelligence (Russell & Norvig, 2022)—noting, however, that the notion of agency in that framework can be rather limited and need not live up to our philosophical explication of agency from Chap. 2. In AI, an agent is defined simply as an entity that is situated in an environment on which it can

¹ In order for the resulting movie to look sensible, there may be certain restrictions on how the shots can be concatenated, or it may be necessary to blur or otherwise process the transitions so that they do not look too disruptive. We can put these worries aside for the moment, and we can imagine the device to be fairly unconstrained, like in a dream, where our brain combines and mixes scenes in sometimes strange ways.

² What we have just described in the context of movies could of course be imagined in a similar way for other media—for example, for music, text, or drawings. In these cases, shots would correspond to bites of sounds, words or syllables, or strokes of the pen or brush.

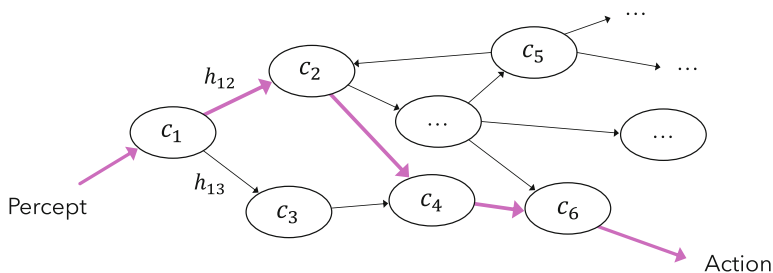


Fig. 5.1 Episodic and compositional memory (ECM) of a PS agent, represented as a mathematical graph. Vertices, denoted by c_j , represent the units of episodic memory called *clips*; edges represent transitions between different clips; real-valued weights h_{ij} encode unnormalised transition probabilities between clips c_i and c_j . Perceptual input from the environment leads to the activation of some clip in the ECM (here, c_1), which then triggers a random walk through a series of associated memory clips (indicated via the pink, thick line), until the walk couples out from another clip (here, c_6), leading to a real action on the environment. Adapted from Briegel and De las Cuevas (2012)

act—thereby possibly changing it—and from which it receives sensory input in the form of percepts.³ Whenever the agent receives a new percept, it needs to find out (or, as one could say, *decide*) what to do next. This process of decision-finding is usually described by an *agent program*. Depending on the type of program, one can distinguish different types of agents. In the AI literature, these types range from simple so-called reflex agents—for example, thermostats or simple versions of robotic vacuum cleaners—to search agents and on to more sophisticated learning and planning agents (see Russell & Norvig, 2022). Further categories of agents can also be defined. To repeat, most of these systems are not agents in the philosophical sense in which we discussed that notion in Chap. 2, but the usage is firmly ingrained in the AI literature.

PS, as a model of agency in the philosophical sense, goes substantially beyond the simple agent-environment model through its sophisticated memory architecture. The relevant key feature of the PS agent is its *episodic and compositional memory* (ECM), which plays a central role for the agent’s capacities to deliberate, to make decisions, and to learn (see Fig. 5.1). The ECM is organised as a directed network of so-called *clips*, comparable to the shots in our description of the random movie player above. Clips are the units of episodic memory. In the simplest form, they are representations of remembered percepts or actions, or short sequences thereof, but they may also have more abstract content. Both the network topology and weights associated with the connections in the network evolve in time via the PS learning

³ In AI and specifically in reinforcement learning (the topic of Sect. 5.3), the environment typically comes with the notion of a *task* that an agent has to solve and is therefore called a *task environment*. While such a view substantially narrows down the general idea of an environment, it is not uncommon—similar views of an environment are also used in biology. In order to start somewhere, we also adopt a minimal notion of environment.

scheme, thereby adapting the agent's behaviour to its environment through a form of reinforcement learning.

Episodic memory is known from psychology (Tulving, 1972) as well as from our daily mental experience; it is the type of memory which stores scenes or experiences from our individual past. Episodic memory is invoked, for example, when we recall what happened when we went to the dentist yesterday, when we first met some person, and so on. Clips in PS can be seen as snippets of episodic scenes. In the context of video or audio processing mentioned above, a clip would describe a short sequence of images or a brief sound bite, respectively. But clips can also represent lower-level perceptual information, such as snippets of sensori-motor impressions—for example, tactile impressions from stroking the surface of an object. The concept of a clip is thus abstract and flexible in its applications.

In PS, a call of episodic memory is triggered by some perceptual input, which starts a random walk through the clip network during which a specific sequence of clips is excited. Any such random walk re-invokes episodic experiences similar, but not necessarily identical, to ones from the past. Re-invoking past experiences in the ECM is similar to replaying a movie with the random movie player in our story above. The only, albeit important, difference is that the clips in the agent's memory are representations of its own experiences. They are generated by the agent itself and are thus intrinsically entangled with the agent's history.⁴ For the PS agent, a random walk in the ECM constitutes a simple form of a *simulation*, in the sense of recalling experiences similar to ones it made in the past when given the same perceptual input. The PS agent uses this process of simulation to prepare the action it will perform next. The transitions between clips in the ECM are stochastic, but the transition probabilities, which are determined by the weights in the network, change dynamically according to the PS learning scheme.⁵

The number of clips representing an agent's ECM is variable and may change over time, as new clips can be created or deleted. The memory is called *compositional* since new clips can be generated in the memory, either through variation (or mutation) or composition of existing ones, or through the creation of completely new clips without a predefined meaning. These intrinsically-created clips need not have a direct correspondence with previous experience. Such *fictional* clips can nevertheless be part of and influence the deliberation process. If a resulting action is

⁴ Staying in the random movie player picture, we could say that the agent generates its own movies, which it continuously re-plays and modifies in the course of new interactions with its environment.

⁵ From the perspective of standard reinforcement learning, which we will discuss in Sect. 5.3, the details of the deliberation are not of interest. Effectively, on an input-output view, a deliberation amounts to the sampling of one of the available actions for a given percept, that is, a sampling of the agent's policy. For PS and its applications as an agency model, however, the details of the deliberation do matter. For example, the actual random walk in the ECM that leads to an action makes a difference for the PS learning rule. The details of modelling deliberation in PS also allow us to understand how an agent formed a given decision, which is important when we connect the model to the philosophical agency and freedom debate. The framework of standard reinforcement learning is therefore too restricted for our discussion.

rewarded, these clips will become integrated into the memory and thus can serve as a basis for the generation of new behavioural patterns. Rules of clip composition and creation can be aligned with the physical processes constituting the agent's memory, and they can be deterministic or indeterministic. Importantly, however, the mentioned dynamics of clips need not be based directly on the agent's external experience and its interaction with the environment (Briegel, 2012).

In mathematical terms, the ECM can be described as a weighted graph. The nodes (clips) of the graph normally carry semantic information, corresponding as it were to different states of mind of the PS agent. The edges (connections) of the graph have real-valued weights encoding transition probabilities between the clips. The dynamics of this graphical model has two components. First, there is a short-term dynamics of deliberation, which is described as a random walk on the given graph, with transition probabilities specified by the weights. Any deliberation thereby corresponds to a certain path on the graph, connecting perceptual input with action output through a sequence of transitions (see Fig. 5.1). Second, there is a long-term dynamics of learning, which changes the transition weights of the graph, and which may also include changes of its topology by adding or deleting nodes as mentioned above. A mathematical description of the dynamics of deliberation and learning will be given in Sect. 5.2.

The terms *projection* and *simulation* highlight two key features of agency already mentioned in Chap. 1, which will be elaborated on in Chap. 6. Projection refers to an agent's capacity to place a given situation in the context of its past experience and to consider different courses of action. The basic meaning of simulation is to go through something similar. In PS, the notion of simulation is used in this basic sense, in that the process of decision finding goes through similar clip sequences as those experienced in the past. These are episodic sequences that, starting with the same perceptual impression, may vary in terms of clip order and content. The functional role of this simulation, realised as a random walk, is thus to re-invoke episodic memories connected with the given perceptual situation, thereby projecting the agent's past experiences into its open future.

The agent's capacities of projection and simulation are based on its memory structure and on the way in which memories are re-invoked by random walks. These capacities thus pertain to the short-term dynamics of deliberation. Deliberation, in turn, has its structural basis in the long-term dynamics of learning. Intimately connected with projection, simulation, and learning is the agent's capacity of modelling. This feature of agency is more advanced, but it is particularly relevant for scientific agency. We will come back to modelling in Sect. 5.8.1.

5.2 Technical Description of the Model

As already mentioned, the PS model adopts the agent-environment framework as used in the fields of artificial intelligence (Russell & Norvig, 2022) and, more specifically, in reinforcement learning (Sutton & Barto, 2018). In that framework,

an agent is always defined relative to an environment from which it receives sensory input in the form of percepts and on which it can act, thereby possibly changing it.

The basic sets (or spaces) of the PS model are the percept space S , the action space A , and the clip space C . The first two sets provide the basic specification of the agent and its environment, describing what the agent can perceive (S) and what actions it can take (A). The third set (C) describes the internal states of memory (clips) that the agent can go through in deliberation. We will assume that S , A , and C are discrete and finite sets, even though this is largely a practical restriction. The percept, action, and clip spaces may have additional structure that reflect the physical constitution of the agent, e.g., its sensory apparatus and motor abilities. To accommodate this possibility, we use the following definitions.

- Percepts are represented as N -tuples $s = (s_1, \dots, s_N) \in S \equiv S_1 \times \dots \times S_N$, $s_i \in \{1, \dots, |S_i|\}$, where the number of possible percepts is given by $|S| = |S_1| \dots |S_N|$. The structure of the percept space S , as a Cartesian product of sets, reflects the different percept categories under which objects can be perceived by the agent. These categories are typically related to the sensory apparatus of the agent. For example, s_1 could label different shapes, s_2 different colors, s_3 different sounds, etc., under which an object can be perceived.
- Actions are represented as M -tuples $a = (a_1, \dots, a_M) \in A \equiv A_1 \times \dots \times A_M$, $a_i \in \{1, \dots, |A_i|\}$, where the number of possible actions is given by $|A| = |A_1| \dots |A_M|$. The structure of the action space A reflects the categories (or degrees of freedom) of the agent's actions. For example $a_1, a_2, a_3, \dots$, could label the states of different motors of a robot, torque levels of its joints, the on/off state of a beeper, etc., depending on the concrete specification of the agent and its environment.
- The action and percept spaces, which constitute the agent's external interface at any given time, may depend on the specific (learning) history of the agent up to that time. In particular, the set of possible percepts that can be encountered by the agent at a given time may depend on the action taken by the agent just before that time and *vice versa*.⁶ We will refine the notation to include such dependencies when needed.

The description of the temporal evolution of the agent involves two distinct dynamical processes. The first process describes deliberation, while the second describes learning. For simplicity, we will assume that these processes happen on different time scales and can be separated. We will further assume, as it is usually done in reinforcement learning, that learning proceeds in discrete steps, or cycles, labelled by an integer t . A cycle then begins with the agent encountering a percept $s \in S$, which triggers a deliberation process that ends with the agent performing some action $a \in A$. In a learning scenario, the agent may also receive further feedback from the environment in the form of a reward R , which then completes

⁶ These mutual dependencies and the (non-quantum mechanical) entanglement of perception and action are also emphasised in the literature on interactive perception (Bohg et al., 2017).

the cycle. The reward typically depends on s and a , but it may also depend, more generally, on the state of the environment and of the agent at the time of the action, or even at other times.

From an external perspective, the history of the agent is given by the record of its interactions with the environment, that is, by the sequence of exchanges of percepts, actions, and rewards. Any characterisation of the agent's behaviour, in particular its capacity to learn, must be based on its history. For a special but important class of learning problems—those that are mathematically well-characterised by so-called Markov Decision Processes (MDPs)—the agent's behaviour can be summarised by a policy function $P^{(t)}(a|s)$. It describes the conditional probability that the agent will perform action a when encountering percept s in cycle t . The policy at time step t is determined by the agent's interaction history until t , and a signature (a measure) of learning is that (how fast) the agent adapts its behaviour over time. In MDPs, such adaptation should be towards an optimal policy $P^*(a|s)$, where a policy is called optimal if it maximises the long-term reward, or expected return (see Sect. 5.3).

While an external perspective on the agent as offered by MDPs is sufficient to discuss its learning performance, a complete account of an agent's history should also include the *history of its deliberations* and the way in which its internal memory changed over time. Adding such an internal perspective will allow us to better understand why, given its past interaction history, an agent behaves in a specific way.

In the PS model, we are particularly interested in the interplay between the internal dynamics of deliberation and the external dynamics of learning. A learning cycle in PS begins with the agent encountering a percept $s \in S$, which leads to the activation of a certain state of its episodic memory (ECM), modelled by some clip $c \in C$. The activation is probabilistic and described by the conditional probability function $I(c|s)$ (“input-coupler”). The excitation of clip c then triggers a deliberation process, which is modelled as a random walk $c \rightarrow c' \rightarrow c'' \rightarrow \dots$ through the set C . For every state c encountered during this random walk, the agent either continues deliberation, or it couples out into an action $a \in A$, a process which is again described by a conditional probability function $O(a|c)$ (“output-coupler”), and which completes the deliberation.

In the following, we will give a more detailed description of the agent's memory, the processes of deliberation, and of learning.

5.2.1 Structure of Episodic Memory

Clips Clips $c \in C$ are the units of episodic memory. In the simplest case, a clip corresponds to a remembered percept or action. Formally, a clip is then a representation $\mu(s)$, or $\mu(a)$, of a single percept s , or a single action a , respectively, in the agent's memory. In this case, we call the corresponding clip *percept clip* $c = \mu(s)$, or *action clip* $c = \mu(a)$, respectively. More generally, clips can be

concatenations of percept clips and action clips, in which case we describe them as L -tuples $c = (\mu_1, \mu_2, \dots, \mu_L)$, where $\mu_l \in \mu(S) \cup \mu(A)$ for $l = 1, \dots, L$. Such clips represent the memories of more extended episodes of the agent's past. Finally, clips may also represent more abstract states of the agent's memory that need not be directly linked to remembered percepts and actions.

Episodic Memory Episodic memory is realized as a stochastic network of clips. Formally, this network is described as a directed weighted graph $G = (V, E)$ where V is a set of vertices and $E \subseteq V \times V$ a set of edges. Every vertex $i \in V$ in the graph G corresponds to a clip c_i . Every edge $(i, j) \in E$ is labelled by a set of weights w_{ij} . The weights that we consider in this Chapter are h -values, related to the hopping or transition probabilities between different clips, and g -values, related to the glow or susceptibility of enhancement of the h -values under reward. Both h and g are real-valued quantities. Further variables may be assigned to an edge, for example emotion tags, or e -values, as introduced in Briegel and De las Cuevas (2012). The edges of the network, together with their weights, may be initialized in arbitrary ways, depending on the type of agent and the situation one considers. In basic two-layered networks, for example, edges typically connect every percept-clip with every action-clip. The h -values of edges (i, j) are initialized to $h_{ij}^{(0)} = 1$, while the g -values for round 0 are only set at the end of that round (see below). The h - and g -values obey $h_{ij}^{(t)} \equiv h^{(t)}(c_i, c_j) \geq 1$ and $g_{ij}^{(t)} \equiv g^{(t)}(c_i, c_j) \geq 0$ at all times t .

5.2.2 Dynamics of Deliberation

Deliberation is understood as a random walk through the clip network. In physical terms, one can think of a physical excitation that hops from clip to clip.⁷ In the following, we describe this random walk (or hopping process) in more detail.

Probability Rule The probability with which an excitation hops from clip c_i to clip c_j is given by the weight $h_{ij}^{(t)}$, normalized by the weights of all edges that connect clip c_i with its neighbours. That is,

$$p^{(t)}(c_j|c_i) = \frac{h^{(t)}(c_i, c_j)}{\sum_k h^{(t)}(c_i, c_k)}, \quad (5.1)$$

where the sum in the denominator is over all outgoing edges connecting clip c_i with some other clip c_k .⁸

⁷ Deliberation is thus structurally similar to remembering, which is simply *re-exciting* clips.

⁸ Other choices for the probability rule are possible and include, for example, a *softmax* version, where $p^{(t)}(c_j|c_i) = \frac{\exp(\beta h^{(t)}(c_i, c_j))}{\sum_k \exp(\beta h^{(t)}(c_i, c_k))}$, which allows for an exponential amplification of sensitivity with the help of some scaling factor β (which plays a role similar to the inverse of

Random Walk The deliberation builds upon this probability rule as follows. As mentioned earlier, the functions $I(c|s)$ and $O(a|c)$, which describe the interface between the PS agent and its external environment, connect the internal random walk, described by the hopping probabilities $p^{(t)}(c_j|c_i)$, with the external behaviour of the agent—its policy—described by $P^{(t)}(a|s)$. The probability that percept s will trigger a specific random walk (of length n) through the clip-network $c_{k_0} \rightarrow c_{k_1} \rightarrow \dots \rightarrow c_{k_n}$ that will eventually lead to action a , is given by $O(a|c_{k_n}) \left\{ \prod_{j=0}^{n-1} p^{(t)}(c_{k_{j+1}}|c_{k_j}) \right\} I(c_{k_0}|s)$. The policy $P^{(t)}(a|s)$ is then given by summing up the probabilities for all paths through the clip network that connect s with a :

$$P^{(t)}(a|s) = \sum_{n=1}^{\infty} \sum_{\pi_G(n)} O(a|c_{k_n}) \left\{ \prod_{j=0}^{n-1} p^{(t)}(c_{k_{j+1}}|c_{k_j}) \right\} I(c_{k_0}|s), \quad (5.2)$$

where the second sum is over all paths $\pi_G(n) = (c_{k_0}, c_{k_1}, \dots, c_{k_n})$ in the graph G , for which $(k_i, k_{i+1}) \in E$ holds for all $i = 0, \dots, n-1$.

Emoticons Optionally, one can assign additional variables $e^{(t)}(c_i, c_j)$ to edges in the clip network, which can influence the deliberation process. In two-layered clip networks, for example, the values of these variables (“emoticons”) may indicate whether the corresponding transition led to a rewarded action the last time it was taken, and may thus contain information about recent changes in the environment. In the simplest case, emoticons are binary variables $e^{(t)}(c_i, c_j) \in \{0, 1\}$ that can be used to inhibit the out-coupling of an excited action clip and re-route the deliberation process, introducing a simple mechanism of reflection, as we will show in Sect. 5.2.4.

5.2.3 Dynamics of Learning

Learning is a process through which the agent adapts its behaviour to the environment. One can imagine that certain types of behaviour are more beneficial to the agent and contribute more to its well-being than others. This could be communicated through a kind of feedback from the environment, which indicates to the agent how well it fits into the environment. In reinforcement learning, one usually models this feedback by a reward function $R = R(s, a)$ that assigns a numerical value to every percept-action pair.⁹

the temperature in a Boltzmann distribution). Softmax is also a convenient way of dealing with negative h -values, in case this generalisation is needed.

⁹ In task environments described by MDPs, the reward is really defined as a function from $S \times A \times S$ to the real numbers, i.e., $R = R(s, a, s')$; see Sect. 5.3.

In the PS model, the reward appears as a parameter in the learning rule, which describes how the state of the episodic memory (specified by the weighted directed graph introduced in Sect. 5.2.1) changes over time. In the first place, there are changes of the weights associated with each of the edges. The learning rule thus refers to the h -values, which directly determine the agent's policy through Eqs. (5.1) and (5.2). To introduce the learning rule, let us assume that, in cycle t , the agent encountered percept $s = s^{(t)}$ and then took action $a = a^{(t)}$, resulting from a deliberation process that involved the specific sequence of clips, or *deliberation path* through ECM,

$$\mu(s^{(t)}) = c_{k_0}^{(t)} \rightarrow c_{k_1}^{(t)} \rightarrow \dots \rightarrow c_{k_n}^{(t)} = \mu(a^{(t)}). \quad (5.3)$$

The *learning rule* can then be stated as follows¹⁰

$$h^{(t+1)}(c_i, c_j) = h^{(t)}(c_i, c_j) - \gamma \left(h^{(t)}(c_i, c_j) - 1 \right) + R^{(t)} g^{(t)}(c_i, c_j). \quad (5.5)$$

Here, $R^{(t)}$ is the numerical reward received by the agent *at the end* of cycle t , while $\gamma \in [0, 1]$ is a damping factor that models forgetting as a basic feature of the PS agent.¹¹ Finally, $g^{(t)}(c_i, c_j)$ —the *glow status*—indicates the susceptibility of the edge (i, j) for the reward that is obtained at the end of round t .

The reward $R^{(t)} \equiv R(s^{(t)}, a^{(t)})$ depends on the percept $s^{(t)}$ that was actually encountered at the beginning of cycle t and on the action $a^{(t)}$ that was actually taken in cycle t . The glow $g^{(t)}(c_i, c_j)$ of edge (i, j) depends on the actual path of deliberation (Eq. (5.3)) the agent went through in cycle t . If the edge (i, j) belongs to the deliberation path taken in cycle t , then its glow is set to its maximal value, $g^{(t)}(c_i, c_j) = 1$. In the beginning (cycle $t = 0$), the glow is thus initialised to 1 for the edges actually traversed, while it is initialised to 0 for all other edges. In later cycles ($t > 0$), the glow of the non-traversed edges is damped according to

$$g^{(t)}(c_i, c_j) = (1 - \eta) g^{(t-1)}(c_i, c_j), \quad (5.6)$$

where $\eta \in [0, 1]$. According to Eq. (5.5), the glow value $g^{(t)}(c_i, c_j)$ describes how susceptible the edge (i, j) is for the enhancement of its h -value if a reward is obtained at time t . A non-vanishing glow value means that a reward also enhances

¹⁰ Sometimes we use a simpler notation and state the learning rule as

$$h_{ij}^{(t+1)} = h_{ij}^{(t)} - \gamma \left(h_{ij}^{(t)} - 1 \right) + R^{(t)} g_{ij}^{(t)}. \quad (5.4)$$

Strictly speaking, h only depends on the edge indices, so h_{ij} suffices. The longer notation $h(c_i, c_j)$ or $h(\mu(s), c_k)$, etc., allows us, however, to specify what type of clips the indices refer to, which sometimes enhances clarity.

¹¹ Note that other conventions for the time labelling of the reward, e.g., $R^{(t+1)} \equiv R(s^{(t)}, a^{(t)})$, are also used in the PS literature and in reinforcement learning (Sect. 5.3), but this is immaterial.

deliberation paths the agent went through some time before t . This enables the agent to deal with delayed rewards.¹²

Similar to the parameter γ , which damps the h -values of all edges, the parameter η describes an exponential damping of glow. Overall, going from round t to round $t + 1$, the h -value of a certain edge in the clip networks grows exactly if the gain term, which is proportional to $R^{(t)}$, is larger than the loss term, which is proportional to γ .

The changes of the weights $h_{ij}^{(t)}$ are due to feedback from the environment; they describe one aspect of learning. Another aspect of learning is clip composition, which may also lead to changes of the topology of the network. Clip composition describes the combination of one or more existing clips into a new clip, which includes the deletion of existing clips and the creation of new ones. Apart from certain syntactic constraints (Briegel & De las Cuevas, 2012) there is much freedom to define the precise rules of clip composition, which then may complement the learning rule for the weights (Eq. (5.5)). Examples will be given in Sect. 5.4.

5.2.4 Basic Examples

To illustrate the dynamics of learning and deliberation, the role of the ECM, and the different learning parameters in Eq. (5.5), let us first look at a few basic examples. In later sections we will connect these examples to more advanced applications in quantum information science and in other fields of research.

Learning and Forgetting A very simple example is that of a binary task environment, where the percept space consists of two elements $S = \{\Leftarrow, \Rightarrow\}$ (perceptual input), and so does the action space $A = \{L, R\}$ (motor output). This describes a scenario where the agent encounters two possible signs in its environment, on which it is supposed to (re-)act in a certain way.

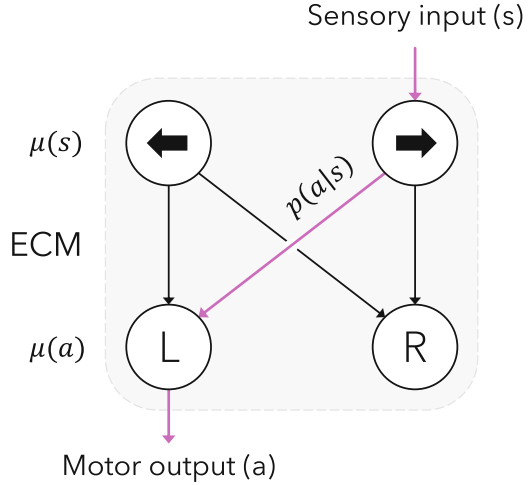
As a concrete task environment, one can imagine car driving, where $S = \{\Leftarrow, \Rightarrow\}$ indicate the possible directions which the car has to follow and $A = \{L, R\}$ the corresponding steering actions, which the driver has to learn to apply correctly. Another example of a task environment is the invasion game studied by Briegel and De las Cuevas (2012), where the agent is supposed to attend to different doors in order to block an invader from entering. In all of these cases, the agent has to learn the *meaning of the symbols* that indicate what it is supposed to do, i.e., which of the different actions will be rewarded.¹³ A reward is issued after each action, i.e., there is no delayed reward.

Figure 5.2 shows the two-layered ECM corresponding to this task environment, where one can see two percept clips, $\mu(\Leftarrow)$, $\mu(\Rightarrow)$ and two action clips, $\mu(L)$,

¹² The glow values g play no role for the *deliberation* process as such, i.e., they do not enter Eq. (5.1), but they influence how a reward is distributed in the ECM and thus how the agent *learns*.

¹³ In the literature, such a scenario is referred to as associative reinforcement learning (Kaelbling, 1994) or as a contextual bandit task (Langford & Zhang, 2007; Sutton & Barto, 2018).

Fig. 5.2 Clip network for the invasion game and similar binary learning tasks. Sensory input s activates a certain percept clip $\mu(s)$, which subsequently makes a transition to one of the action clips $\mu(a)$ with probability $p(a|s) \equiv p(\mu(a)|\mu(s))$, as given by Eq. (5.8). Adapted from Briegel and De las Cuevas (2012)



$\mu(R)$. The clips $\mu(s)$ and $\mu(a)$ are graphically represented by circles and are to be distinguished from the external (“raw”) sensory input s and motor output a , respectively. In this case the learning rule has the form

$$h^{(t+1)}(\mu(s), \mu(a)) = h^{(t)}(\mu(s), \mu(a)) - \gamma \left(h^{(t)}(\mu(s), \mu(a)) - 1 \right) + R^{(t)}, \quad (5.7)$$

and the clip-to-clip transition probabilities are, for example,

$$p^{(t)}(\mu(R)|\mu(\Rightarrow)) = \frac{h^{(t)}(\mu(\Rightarrow), \mu(R))}{h^{(t)}(\mu(\Rightarrow), \mu(R)) + h^{(t)}(\mu(\Rightarrow), \mu(L))}. \quad (5.8)$$

If one assumes deterministic input- and output couplers and a one-to-one correspondence between external sensory input and motor output and their internal representations, i.e., $I(c|s) = \delta(c, \mu(s))$ and $O(a|c) = \delta(\mu(a), c)$, one obtains for the policy

$$P^{(t)}(a|s) = p^{(t)}(\mu(a)|\mu(s)). \quad (5.9)$$

For notational ease we often write $h^{(t)}(\mu(s), \mu(a)) \equiv h^{(t)}(s, a)$ and $p^{(t)}(\mu(a)|\mu(s)) \equiv p^{(t)}(a|s)$, where it is understood that the arguments of h and p refer to the internal representations of s and a , respectively.

Equation (5.7) can be solved numerically to compute, for example, the learning curves as shown in Fig. 5.3.¹⁴

¹⁴ For the initial slope and the asymptotic values of the learning curves, one can even derive (approximate) analytic solutions (Mautner et al., 2015).

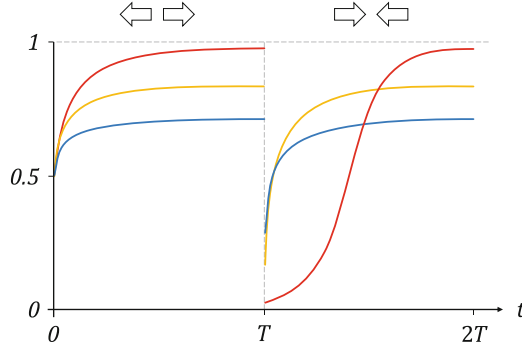


Fig. 5.3 Schematic learning curves for a basic PS agent in the invasion game. Symbols $\Rightarrow$ and $\Leftarrow$ are encountered randomly. The agent learns the meaning of the symbols typically after a few tens of steps $t \simeq T$. Different values of γ in Eq. (5.7), indicated by different colours of the learning curves (*red*: small, *yellow*: medium, *blue*: large), result in different values of the reachable blocking efficiency (highest for *red*, lowest for *blue*). This efficiency indicates how well the correct action for the two different symbols has been learnt. It is quantified by the (asymptotic value of) the average reward $r^{(t)}$. At time step $t = T$, the meaning of the symbols is inverted; the symbol $\Leftarrow$ ($\Rightarrow$) now indicates that the agent is supposed to go right (left). Since, at this stage, the agent has already memorised the previous meaning of the symbols, it needs some time to adapt to the new situation. While a smaller damping factor leads to a higher blocking efficiency, it also takes the agent a longer time to forget the old and learn the new meaning of the symbols. Adapted from Briegel and De las Cuevas (2012)

The learning curve describes the evolution of the average reward $r^{(t)}$, for $R^{(t)} \in \{0, 1\}$, as a function of the time step t :

$$\begin{aligned} r^{(t)} &= \sum_s P^{(t)}(s) \sum_a R^{(t)} P^{(t)}(a|s) = \sum_s P^{(t)}(s) P^{(t)}(a_s^*|s) \\ &= \sum_s P^{(t)}(s) p^{(t)}(a_s^*|s), \end{aligned} \quad (5.10)$$

where a_s^* denotes the correct (i.e., rewarded) action for symbol s . The third equality is due to Eq. (5.9). For a task environment that generates percepts randomly, we choose $P^{(t)}(s) = 1/2$. At time step T , the meaning of the symbols is inverted, i.e. the symbol $\Leftarrow$ ($\Rightarrow$) now indicates that the attacker is going to move to the right (left). Since the agent has already built up some memory, it needs some time to adapt to the new situation. For changing environments like this, forgetting can be helpful, which is modelled by a damping factor $\gamma > 0$. In Fig. 5.3 one can see a trade-off between adaptation speed, on the one side, and achievable blocking efficiency, on the other side. This task environment can easily be generalised and solved for larger percept and action spaces.

Reflection In Fig. 5.2, we highlighted the difference between percepts and actions (i.e., external sensory input and motor output), on the one hand, and their internal representations as percept clips and action clips, on the other hand.

The virtue of this distinction becomes apparent if we consider deeper clip networks, as we will do shortly. In these applications the focus lies on the deliberation rather than the learning dynamics, and PS is also used, e.g., as a model for reasoning in the sense of decision finding. But even in two-layered networks the distinction plays a role if we consider the possibility of *reflection* (Briegel & De las Cuevas, 2012). This can be realised, for example, by the use of emoticons $e_{sa}^{(t)} \equiv e^{(t)}(\mu(s), \mu(a))$ as additional weights associated with the edges $(\mu(s), \mu(a))$ of the clip network in Fig. 5.2. We consider the simplest case, where these weights are binary variables. Whenever a transition $(\mu(s) \rightarrow \mu(a))$ is taken during the deliberation, the emoticon $e_{sa}^{(t)}$ is set to the value 1 when the resulting action is rewarded and it is set to 0 otherwise. The emoticon remembers, as it were, if the edge was part of a successful deliberation the last time it was taken. This provides a simple way of conditioning the termination of deliberation and its coupling out into real action. To see this, we assume that the excitation of an action clip $\mu(a)$ as in Fig. 5.2 triggers an action only if the emoticon on the transition leading to the clip $\mu(a)$ has the value 1, while otherwise the random walk is reset and starts with the excitation of the same initial clip. This allows the agent to reconsider, as it were, its choice. The maximum number of random walks per decision can be limited by a predefined reflection-time parameter N_R , whose default value $N_R = 0$ means no reflection. Reflection is illustrated in Fig. 5.4. In this case, the probability for the

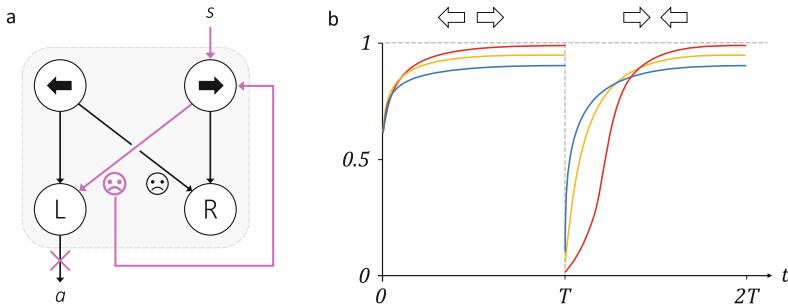


Fig. 5.4 PS with reflection. **(a)** Clip network for the invasion game including reflection. Binary flags (emoticons) on the clip-to-clip transitions, realised as binary weights of the corresponding edges of the ECM, can be used to inhibit the out-coupling of a memory excitation into a real action. In the figure, an excitation of action clip $\mu(L)$ does not lead to the corresponding motor action $a = L$ but to a re-excitation of clip $\mu(\Rightarrow)$. Out-coupling of $\mu(L)$ is inhibited due to a negative flag on the transition $\mu(\Rightarrow) \rightarrow \mu(L)$. The flag could, for example, encode a negative experience associated with action $a = L$, given the percept $s = \Rightarrow$. **(b)** Effect of reflection on the achievable blocking efficiency. Already a single round of reflection, described by $N_R = 1$ in Eq. (5.11), can significantly increase the achievable blocking efficiency. The figure shows the learning curves of Fig. 5.3, albeit for $N_R = 1$. Adapted from Briegel and De las Cuevas (2012)

agent to take the correct action a_s^* when encountering symbol s is given by

$$P^{(t)}(a_s^*|s) = 1 - \left(1 - p^{(t)}(\mu(a_s^*)|\mu(s))\right)^{N_R+1}, \quad (5.11)$$

which replaces Eq. (5.9) and leads to a corresponding modification of the average reward (Eq. (5.10)). Figure 5.4b shows the learning curves of Fig. 5.3 for $N_R = 1$. Comparing this to the learning curves shown in Fig. 5.3, one can see that the agent achieves much higher blocking efficiencies, even with significant values of damping.

Deliberation and Decision Finding The example described in the previous paragraph illustrates a basic feature of PS agents, namely, that they can reflect or, more generally, deliberate in a meaningful way. A given input does not need to lead to an action directly. Deliberation does not have to involve reflection and emoticons as above, but can also be modelled in plain ECMs once these have more than two layers. Such deeper structures of PS help to resolve the process of deliberation or decision-finding itself—for example, if we want to understand precisely how a decision was made by an agent. This is relevant for philosophical applications of PS as an agency model, but also for the field of explainable AI and for the study of artificial agency more generally.

If we want to resolve the deliberation process, it is often useful to think of the clips as *considerations*, and to use PS as a basic model of reasoning. The random walk in the ECM is then understood as a progression of associative considerations triggered by a given perceptual situation. In Müller and Briegel (2018), we discuss a toy model for the deliberation of an agent that is about to take an action, using the clip network shown in Fig. 5.5. The agent we consider is a cat who sees a sunny spot on his owner’s desk and has to decide whether to jump there or to walk away. The network has a single percept clip $\mu(s)$ (seeing a sunny spot on the desk) as a starting state, and two action clips $\mu(a_1)$, $\mu(a_2)$ that, when activated, will trigger two different actions a_1 (jumping on the desk) and a_2 (walking away), respectively. The other three intermediate clips of the network, c_1 , c_2 , and c_3 , represent remembered episodes that the cat experienced in similar situations in the past—something like the cosy feeling of the warm desk, the owner’s attention, and the owner’s feeling disturbed and brushing the cat away, respectively. The successive excitation of these intermediate clips can be seen as a progression of the cat’s considerations in the given situation, leading eventually to one of the actions a_1 or a_2 . In this example, clips encode low-level considerations such as the mentioned sensori-motor impressions of the cat. The random walk in the ECM is accordingly interpreted as a progression of such considerations, triggered by some perceptual situation and eventually leading to some action. Mapping out a concrete deliberation path through the network then provides a fairly low-level, but still reasons-type explanation of how the cat came to a specific decision. Furthermore, for a given state of the agent’s memory, specified by all h -values,¹⁵ one can precisely compute the probabilities

¹⁵ As mentioned in Footnote 12, the glow values g play no role for the deliberation process.

for all possible deliberation paths that begin with percept s and end with one of the actions a_1 or a_2 , and other statistical quantities such as the average deliberation time. For example, the probability that the cat will eventually jump (action a_1) is given by Eq. (5.2) with $I(c_{k_0}|s) = \delta(c_{k_0}, \mu(s))$, $O(a_1|c_{k_n}) = \delta(\mu(a_1), c_{k_n})$, and thus with the summation running over all paths that begin with $c_{k_0} = \mu(s)$ and end with $c_{k_n} = \mu(a_1)$. Writing this out, we obtain

$$\begin{aligned} P^{(t)}(a_1|s) &= p^{(t)}(\mu(a_1)|c_1)p^{(t)}(c_1|\mu(s)) \\ &+ p^{(t)}(\mu(a_1)|c_2)p^{(t)}(c_2|c_1)p^{(t)}(c_1|\mu(s)) \\ &+ p^{(t)}(\mu(a_1)|c_1)p^{(t)}(c_1|c_3)p^{(t)}(c_3|c_2)p^{(t)}(c_2|c_1)p^{(t)}(c_1|\mu(s)) + \dots \end{aligned} \quad (5.12)$$

For illustration, the probabilities of some of the deliberation paths are shown in Table 5.1 for the special case when all transitions from any given clip have the same weight. In this case, $p^{(t)}(c_1|\mu(s)) = 1$, and $p(\cdot|\cdot) = \frac{1}{2}$ for all the other transitions in Fig. 5.5. The probability that the cat eventually jumps then sums up to $P^{(t)}(a_1|s) = 6/7$.

In Müller and Briegel (2018) this detailed and quantitative account of the stochastic deliberation dynamics was used in the context of a philosophical discussion of free agency and free will under indeterminism. One of the main issues in that area is the so-called luck objection, according to which indeterminism is inimical to freedom: it is argued that indeterminism in the process of decision finding means that a would-be agent is in fact subject to pure chance, so that no meaningful decisions can occur. On the contrary, we hold that the PS model's decision dynamics, together with the longer-term dynamics of learning, can successfully anchor meaningful decisions based on indeterminism. In the mentioned paper we make this point specifically with reference to Van Inwagen's replay argument, which is a much-discussed variant of the luck objection (Müller & Briegel, 2018, §4.3).

Clips need not be directly tied to sensori-motor impressions as in the example of the cat who decides whether to jump on the desk or to walk away. Even in examples that are structurally similarly simple, clips can also represent higher-level

Table 5.1 Some deliberation paths in the ECM of Fig. 5.5, together with their corresponding probabilities as obtained from Eq. (5.12)

Path	Length l	Probability $\Pr(l)$	Jump (a_1) or walk away (a_2)?
$\mu(s) \rightarrow c_1 \rightarrow \mu(a_1)$	2	0.5	Jump
$\mu(s) \rightarrow c_1 \rightarrow c_2 \rightarrow \mu(a_1)$	3	0.25	Jump
$\mu(s) \rightarrow c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow \mu(a_2)$	4	0.125	Walk away
$\mu(s) \rightarrow c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_1 \rightarrow \mu(a_1)$	5	0.0625	Jump
$\mu(s) \rightarrow c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_1 \rightarrow c_2 \rightarrow \mu(a_1)$	6	0.03125	Jump
$\mu(s) \rightarrow c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow \mu(a_2)$	7	0.015625	Walk away
...	...	...	...

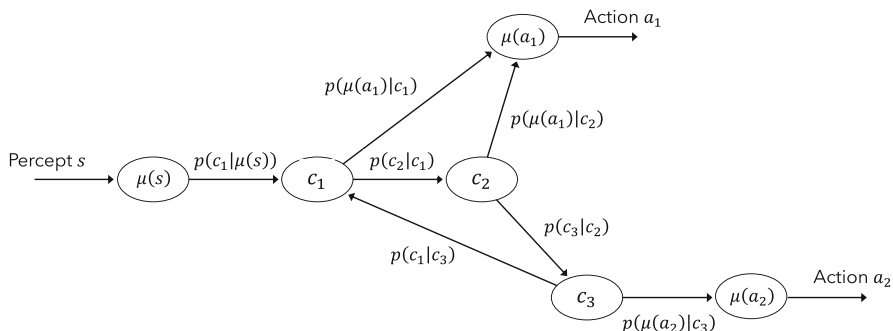


Fig. 5.5 Episodic memory generating a chain of considerations of an agent that is about to take an action. Note that the directed graph describing the ECM exhibits a cycle between a set of clips, that is, a closed path $c_1 \rightarrow c_2 \rightarrow c_3 \rightarrow c_1$. The existence of a cycle, together with the indeterministic nature of the deliberation, modelled as a random walk, can lead to deliberation paths of variable lengths as listed in Table 5.1. These paths describe a basic process of reasoning. Adapted from Müller and Briegel (2018)

considerations. In Chap. 6, where we introduce our phenomenological interpretation of the PS model, we will discuss a relevant example in detail. Consider the situation depicted in Fig. 6.1, in which a human agent considers different uses of a hammer—from driving in a nail to smashing a window to escape a house that is on fire—and selects the one that is relevant in her circumstances. That example comes with a much richer semantic context—like, what it means that a house is on fire, and why the agent had better use the hammer to break the glass of the window and escape from the room, rather than continue working with the hammer on something else. In that example, the semantic context of the considerations gives them their relevance, and the different possible actions in the given situation carry fundamentally different weights.

Learning from Delayed Rewards In many contexts, a reward cannot be obtained immediately after every action, but may be delayed and depend on an entire *sequence* of actions. In PS, the glow mechanism plays a crucial role in modelling delayed rewards. A textbook example for delayed rewards is the task environment called **grid world**. The **grid world** agent can move between different positions, or sites, on a grid.¹⁶ At each step it can choose to take one of four different actions, which correspond to moving left, right, up, or down. One of the positions on the grid, the *target position*, is distinguished by the fact that the agent receives a positive reward exactly when it moves to this position. **Grid world** is usually described as a task environment, where the task of the agent is to find the rewarded target site with

¹⁶ It is usually assumed that the positions have labels or are otherwise identifiable, so that being at different positions means having different percepts.

the smallest number of steps.¹⁷ To measure the learning performance, one looks at the shortest path found by the agent in a given number of trials. Every trial is initialised with the agent starting at the same position; then it moves around until it finds the target site and is rewarded, or it reaches a maximum number of allowed steps without finding the target site. In both cases, it is reset to the starting position and the next trial starts. The main question is how many trials the agent needs to take until it has learnt to find the target in the shortest number of steps.

The problem can be generalised in various ways. For example, there may be obstacles at certain places, which means that one or more of the four basic actions are effectively not available to the agent. If the agent tries to move in the direction of the obstacle, it will typically be set back to its current position. In a similar way, one can set boundaries to the grid, in which case the agent is confined to move within a finite region. One can also reward several sites of the grid, possibly with different rewards, and increase the dimension of the grid, which amounts to increasing the number of actions that are available at any site. The grid can also be replaced by a general graph, with the reward being attached to a specific vertex. In this situation, the size of the action space (the number of available actions) depends on the vertex (and thus, on the percept) and is given by the local degree of the vertex in the graph. This exemplifies the general situation of variable action and/or perception spaces described at the beginning of Sect. 5.2. We will discuss a concrete application with this feature in Sect. 5.4.4.

For a PS agent, the feature of glow enables the agent to solve the *grid world* task and, more generally, to learn from delayed rewards. This can be understood as follows. When the agent finds the rewarded site for the first time, it will have made a sequence of steps (actions) $a^{(0)}, a^{(1)}, \dots, a^{(T-1)}$ that map out a certain (random) path $s^{(0)}, s^{(1)}, \dots, s^{(T)}$ of length T through the grid, from the starting site $s^{(0)}$ to the target site $s^{(T)}$. For the PS agent, a step from site $s^{(t)}$ to some neighbouring site $s^{(t+1)}$, realised by the action $a^{(t)}$ in step t , is preceded (simulated) by a corresponding clip-to-clip transition in the ECM between percept clip $\mu(s^{(t)})$ and action clip $\mu(a^{(t)})$, whose glow status is then set to $g^{(t)}(\mu(s^{(t)}), \mu(a^{(t)})) = 1$. The same happens to all edges $(\mu(s^{(t')}), \mu(a^{(t')}))$ corresponding to transitions that are taken in the ECM as part of the agent's deliberations while walking in the grid. However, while the glow value of each transition that has been freshly taken will be set to unity at that time, the glow on that transition will be reduced by a factor of $1 - \eta$ every time the agent makes another step (as long as it does not make the same transition again, otherwise glow on this edge will be freshly set to unity). At the time $t = T - 1$ when the agent makes its last move to the rewarded site, the edges of the (internal) transitions in the agent's memory $(\mu(s^{(t')}), \mu(a^{(t')}))$, for

¹⁷ This way of speaking makes proper sense for standard reinforcement learning agents, which are designed with a generic goal (or task) being given (see Sect. 5.3). Their goal is to maximise a utility function, viz., the long-term return as defined in Eq. (5.14); see also Footnote 32. Due to the occurrence of the discount factor $\gamma_d < 1$ in Eq. (5.14), the utility is largest if the reward is obtained after the smallest possible number of steps. In the *grid world* environment, this translates into the task of finding the shortest path between the starting and the goal state.

$t' = 0, 1, \dots, T - 1$, will be glowing with different intensity, the weaker the earlier they were taken.¹⁸ The glow thus reflects how many steps ago a transition in the clip network was taken before the agent reached the target.

After the final step, that is, at the end of round $T - 1$, a reward is issued which will increase all h -values in the ECM in proportion to their glow value at that time (see Eq. (5.5)). In this way the reward is effectively distributed backward in time and will also affect transitions that were taken by the agent on its way before reaching the target. In the next trial, the agent will again perform a random walk, starting from the same initial site. This time, actions that were taken at a given site in the previous trial will be taken again with a larger probability, such that the agent is more likely to follow a path that is similar to the one taken last time. However, since the actions at each site are still probabilistic, the agent may also take other actions with certain probabilities and thereby, as it were, explore new paths. If it thereby finds a shorter path, or a shortcut to the previous path, the corresponding transitions in the ECM will be rewarded more strongly. As the trials are repeated many times, the agent will thus find shorter and shorter paths; at the same time, the agent will be more and more likely to stay on these shorter paths, since the probabilities for actions that leave these paths will become progressively smaller. Overall, based on the mechanism of glow, the agent will thus develop a policy that generates close-to-optimal paths.

In Fig. 5.6, we show a two-dimensional grid world with some obstacles and a path the agent learns to take. The direction of the arrows on the path indicates the respective actions with the highest h -values. The intensity of the arrows indicates the value of glow (the deeper the colour, the closer the g -value is to 1). Interestingly, learning the shortest paths requires an intermediate value of the glow damping parameter $0 < \eta < 1$. If $\eta \simeq 1$, the glow on each transition is damped away too quickly, such that the agent receives no guidance at sites that are even a little away from the target site. If $\eta \simeq 0$, high values of glow will persist on all transitions that were taken in the past. For large rewards, this leads to a reinforcement of non-optimal paths while the exploration of shorter paths will be suppressed. For small rewards, different paths will be explored. Consequently, all transitions that are taken in different trials will be rewarded, which may again lead to a flat probability distribution over the available actions at each site.

In Sect. 5.4 we will discuss more advanced task environments that include delayed rewards. In these examples a PS agent is used to control a quantum environment, e.g., in the context of adaptive quantum error correction and in the design of quantum experiments.

¹⁸ If we assume, for simplicity, that the path taken has no loops, then the corresponding glow values would be $g^{(T-1)}(\mu(s^{(t')}), \mu(a^{(t')})) = (1 - \eta)^{T-t'-1}$.

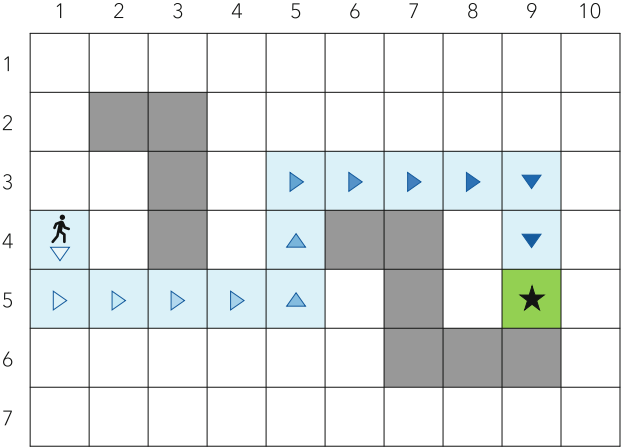


Fig. 5.6 Example of a 2D grid world environment. The agent starts at an initial position that is held fixed throughout the learning task (here, position (1, 4)) and has to find the target (here, at position (9, 5), indicated by a star) while avoiding obstacles (grey filled boxes). A reward is issued only when the agent arrives at the target. The blue fields mark a path that a successful PS agent finds after learning. The arrows in these fields indicate the action that has the highest h -value, and the shading of the arrows indicates their glow status at the end of learning, the darker the closer to 1

5.3 Contrasting PS with Standard Reinforcement Learning

Reinforcement learning denotes one of the three main paradigms of machine learning. The other two are called supervised learning and unsupervised learning, both of which deal with the problem of classifying data into different subsets, or classes. These classes are either learnt from labelled examples (supervised) or they emerge by clustering the input into subsets of data that are similar under some given metric (unsupervised). Reinforcement learning is different from both of these paradigms insofar as the agent itself generates the data while interacting with its environment and learns through these interactions.

The mathematical framework that underlies most of the theory of reinforcement learning is that of Markov Decision Processes (MDPs). MDPs provide a formalism for *sequential decision making* based on the notions of states, actions, and rewards, in a similar way as we introduced them in Sect. 5.2. In addition, MDPs include a description of the *environment* of the agent, in the form of a transition function which specifies the effect of the agent's actions on the state of the environment. Together, we have the following formal specification (see Van Otterlo & Wiering, 2012):

Markov Decision Process (MDP) A Markov Decision Process is a 4-tuple (S, A, T, R) , where S denotes a finite set of states, A a finite set of actions,

$T : S \times A \times S \longrightarrow [0, 1]$ a transition function and $R : S \times A \times S \longrightarrow \mathbb{R}$ a reward function.

The functions T and R together define the environmental model of the MDP. The transition function T specifies the conditional probability that the agent, after taking action $a \in A$ in state $s \in S$, will find the environment in state $s' \in S$:¹⁹

$$\Pr(s'|s, a) = T(s, a, s'). \quad (5.13)$$

Similarly, the reward function R assigns a numerical reward to every state-action-state triplet (s, a, s') : if the environment is transformed to state s' as a result of the agent taking action a in state s , then the reward is $R(s, a, s')$.²⁰ Since the resulting state s' is a random variable, whose probability is specified by the transition function, the reward cannot be fully controlled by the agent even for deterministic reward assignments. The latter is assumed here, following Van Otterlo and Wiering (2012). If the reward assignment itself is probabilistic, i.e., if the reward r is also considered to be a random variable, then $R(s, a, s')$ is understood as the expected reward for the triplet (s, a, s') .²¹ Note that the probability in Eq.(5.13) depends only on the state s in which the action is taken and on the action a itself, but not on earlier states encountered and actions taken by the agent. This is the Markov (“no history”) property of the environment as assumed by an MDP. It is a strong restriction and many real environments do not have this property, but it is still often a useful approximation.

The theory of MDPs deals with the problem of optimal sequential decision making.²² The main objective of the theory is to find an agent’s *optimal policy* in an environment modelled by the MDP. The policy $\pi : S \times A \longrightarrow [0, 1]$ assigns to each state-action pair (s, a) the conditional probability $\pi(a|s)$ that the agent, upon encountering state s , takes action a . The policy is *optimal*, written $\pi = \pi^*$, if it maximises a certain performance measure called *expected return*. Informally, this is the total reward the agent can expect to collect (within a certain time horizon) if it follows the policy under consideration at each time step. Suppose that the application of that policy generates the interaction history $\dots s^{(t-1)}, a^{(t-1)}, R^{(t-1)}$,

¹⁹ For deterministic environments, one often chooses an alternative definition where T maps state-action pairs (s, a) to successor states $s' = s'(s, a)$. In the above definition, the deterministic case can be obtained by the choice of a delta-type transition probability $T(s, a, s') = \delta(s', s'(s, a)) \in \{0, 1\}$.

²⁰ The reward R is sometimes defined, alternatively, as a function that maps state-action pairs (s, a) to real numbers $R(s, a)$. Such a situation can be recovered as a special case of the above definition.

²¹ For such fully indeterministic situations, a more compact and elegant description of MDPs can be given (Sutton & Barto, 2018) by defining the *dynamics* of the MDP as a conditional probability function $p(s', r|s, a)$, from which both the transition function $T(s, a, s') = \sum_r p(s', r|s, a)$ and the (expected) reward $R(s, a, s') = (\sum_r r p(s', r|s, a)) / (\sum_r p(s', r|s, a))$ are obtained.

²² Historically, discrete MDPs were introduced and developed in the context of control theory and dynamic programming (Bellman, 1957; Sutton & Barto, 2018), where one talks of a controller that makes decisions to control the behaviour of a dynamical system, instead of an agent that takes actions in an environment.

$s^{(t)}, a^{(t)}, R^{(t)}, s^{(t+1)}, a^{(t+1)}, R^{(t+1)}, \dots$, which is here considered as a sequence of values taken by the corresponding random variables s, a, R at the different time steps. The return $G^{(t)}$ for that specific interaction history (random sample), from time step t onward, is then defined as the sum

$$G^{(t)} = \sum_{k=0}^{\infty} \gamma_d^k R^{(t+k)}, \quad (5.14)$$

with a discount factor $\gamma_d \in [0, 1]$.²³

With the help of this performance measure it is possible to assign, for a given policy π , to each state $s \in S$ a numerical value

$$V_{\pi}(s) = E_{\pi}[G^{(t)} | s^{(t)} = s], \quad (5.15)$$

where $E_{\pi}[\cdot]$ denotes the expectation value, given that the agent follows policy π . The *state-value function* $V_{\pi}(s)$ thus measures how good it is for the agent to be in state s , in terms of the long-term return it can expect, when following policy π .²⁴ In a similar way one can define an *action-value function*

$$Q_{\pi}(s, a) = E_{\pi}[G^{(t)} | s^{(t)} = s, a^{(t)} = a], \quad (5.16)$$

which measures the expected return if the agent takes action a in state s and thereafter follows policy π .²⁵

Value functions, and utilities in general,²⁶ allow one to define an explicit goal for the learning agent that it is supposed to reach. Informally, the goal of a learning system is to behave optimally in its environment. For MDPs, this means to find, or

²³ Strictly speaking the definition applies to the infinite horizon model of the return. Note that the rewards are discounted the more the further they lie in the future of step t , which defines an effective time horizon of the order of $(1 - \gamma_d)^{-1}$ steps that is determined by γ_d . Any value of the discount factor $\gamma_d < 1$ mathematically guarantees the convergence of the infinite sum. Alternative models exist where the sum is finite or replaced by an average, both cases allowing $\gamma_d = 1$ (no future discounting).

²⁴ Note the difference: “How good is it for the agent to be in state s when following policy π ” (which plays a role for choosing the next action) versus “How good it is for the agent to follow policy π when it is in state s ” (which plays a role for improving the policy).

²⁵ Here it is not presupposed that the agent follows the policy $\pi(a|s)$ when choosing the action a , which appears in the argument of $Q_{\pi}(s, a)$. The probability $\pi(a|s)$ could be vanishingly small.

²⁶ The expected return in Eq. (5.15) can also be seen as measuring some utility (Van Otterlo & Wiering, 2012), and often these terms are used in an interchangeable way. (Depending in the context one may then speak of “the utility of following a certain policy when being in a given state”, or “the utility of being in a certain state when following a given policy”, or “the utility of taking a certain action in a given state when subsequently following a given policy”, etc.) For a more general discussion of the role of utility functions and their maximisation in cognitive and behaviour modelling, see Lewis et al. (2014).

to learn, an optimal policy. With the help of Eq. (5.15), this can be formalized as a policy that simultaneously maximizes the value function $V_\pi(s)$ for every state s .

The mathematical significance of the value function as defined in Eq. (5.15) lies in the fact that it satisfies a set of recursion relations. From Eq. (5.14) together with Eq. (5.15) one sees that $G^{(t)} = R^{(t)} + \gamma_d G^{(t+1)}$, which can be used in the derivation of a so-called *Bellman optimality equation* (Bellman, 1957), a version of which exist also for the action-value function Eq. (5.16). For a given model for the environment—that is, for known functions $T(s, a, s')$ and $R(s, a, s')$ —and for finite MDPs, the optimal value function $V^*(s) = \max_\pi V_\pi(s)$ can be obtained by solving the Bellman optimality equation. Their solution in turn allows one to obtain an optimal policy π^* , which is the policy the agent needs to follow to maximise the expected return (Sutton & Barto, 2018, Ch. 3.6).

For our context it is particularly important that other methods have been developed in the field of reinforcement learning that do not assume explicit knowledge of the environmental model, but instead build upon interactions of the agent with the real environment (Sutton & Barto, 2018; Kaelbling et al., 1996; Van Otterlo & Wiering, 2012). Estimates of the value function can be obtained by trying out different actions, effectively sampling the transition function through Monte-Carlo type methods. *Temporal difference learning* further combines Monte-Carlo methods with iteratively improving the estimates of the optimal value function from round to round, under some update rule. The simplest version of a temporal difference learning rule, which we give here for the action-value function, takes the form (Sutton & Barto, 2018, Ch. 6.4, p. 129)

$$\begin{aligned} Q^{(t+1)}(s^{(t)}, a^{(t)}) &= Q^{(t)}(s^{(t)}, a^{(t)}) \\ &+ \alpha \left[R^{(t)} + \gamma_d Q^{(t)}(s^{(t+1)}, a^{(t+1)}) - Q^{(t)}(s^{(t)}, a^{(t)}) \right], \end{aligned} \quad (5.17)$$

where α is some adjustable learning parameter. Here, $s^{(t)}$ and $a^{(t)}$ refer to the state actually encountered and action actually taken in round t , respectively. Note that the right hand side of Eq. (5.17) includes the term $Q^{(t)}(s^{(t+1)}, a^{(t+1)})$, which refers to a future state $s^{(t+1)}$ and action $a^{(t+1)}$ not yet encountered.

The learning rule given by Eq. (5.17) is used in the popular algorithm SARSA and, with a small modification,²⁷ in Q-learning. The latter algorithm is especially useful for deterministic policies (Watkins & Dayan, 1992): once the optimal action-value function $Q^*(s, a)$ has been obtained, the corresponding policy can be extracted in a straightforward way, $\pi^*(s) = \arg \max_a Q^*(s, a)$. For the more general case of stochastic policies the process is more complicated and involves, for example, policy iteration methods where the update of the value function with respect to a fixed policy alternates with an update of the policy based on the new

²⁷ This modification consists in the replacement of $Q^{(t)}(s^{(t+1)}, a^{(t+1)})$ with the expression $\max_a Q^{(t)}(s^{(t+1)}, a)$ in Eq. (5.17) (Sutton & Barto, 2018, Ch. 6.5, p. 131).

value function. Alternatively, so-called policy gradient methods use (families of) parametrised policies as variational families to directly maximise their associated value function. In any case, value functions such as $V(s)$ and $Q(s, a)$ play a central role in most realisations of reinforcement learning. As is highlighted by Sutton and Barto (2018, 73), the “key idea of [...] reinforcement learning generally, is the use of value functions to organize and structure the search for good policies”.

This approach is different from Projective Simulation, which does not involve the concept of a value function. On first sight, the action-value function $Q(s, a)$ for an MDP and the h -values $h(c_i, c_j)$ for an ECM may look similar, especially for a two-layered ECM. In that case, and without the mechanism of reflection, we may neglect the distinction between states/actions and their internal representations as state/action clips in PS and write $h(c_i, c_j) = h(\mu(s), \mu(a)) = h(s, a)$. The learning rule of Eq. (5.5) then updates all h values according to

$$h^{(t+1)}(s, a) = h^{(t)}(s, a) - \gamma \left(h^{(t)}(s, a) - 1 \right) + R^{(t)} g^{(t)}(s, a). \quad (5.18)$$

For the actual state-action-reward sequence encountered by the agent in round t , i.e., the sequence $s^{(t)}, a^{(t)}, R^{(t)}$, we have $g^{(t)}(s^{(t)}, a^{(t)}) = 1$, so that

$$h^{(t+1)}(s^{(t)}, a^{(t)}) = h^{(t)}(s^{(t)}, a^{(t)}) - \gamma \left(h^{(t)}(s^{(t)}, a^{(t)}) - 1 \right) + R^{(t)}, \quad (5.19)$$

which may be compared with Eq. (5.17). It is tempting to associate the discounting parameter γ_d with the forgetting parameter γ (or, $1 - \gamma$) in Eq. (5.19) and to scale the parameter α such as to make the equations look more similar. This is, however, misguided, since forgetting is a mechanism in PS that works directly on the h -values, while discounting is applied to the (future) rewards in the long-term return G_t of Eq. (5.14), which affects Q internally. But the main difference, which is also the most important one on the technical level, is that in contrast with Eq. (5.17), in the PS learning rule (Eq. (5.19)) there is no time index $t + 1$ appearing anywhere in the arguments on the right hand side. Thus, in order to update $h^{(t)}(s, a)$ on a state-action pair $(s, a) = (s^{(t)}, a^{(t)})$ encountered in round t , there is no need for PS to take into account the next state $s^{(t+1)}$ or action $a^{(t+1)}$ (sampled by using the policy). In other words, the PS learning rule is strictly *local in time*.²⁸

Both reinforcement learning and PS are based on the agent-environment paradigm. The contrast between reinforcement learning and PS can be further strengthened by considering how and at which place the *agent* enters the stage in the development of the frameworks. Reinforcement learning starts from a goal, derived from the idea of a task environment, and then constructs a learning rule such as to satisfy this goal. In contrast, PS starts with a model of episodic memory and a

²⁸ PS therefore does not use temporal difference learning in the proper sense. As discussed later in the text, the glow mechanism does effectively accumulate discounted rewards, thus enhancing connections based on data from the past. Importantly, however, there is never a reference to future times.

simple learning rule as its basic resources, and then exposes the agent to tasks that may be similar as in reinforcement learning. It is, however, not necessarily assumed, and not guaranteed, that the PS agent is capable of learning the optimal behaviour. The PS model was indeed not developed with this objective, but with the objective of providing a plausible and useful model for agency, which includes decision-finding and reasoning, and a form of learning. But this form of learning is not constructed such as to optimise some given utility. PS agents need no utility function.²⁹

When comparing the learning dynamics of PS with standard reinforcement learning, the physical motivation of PS also comes to the fore. The h -values in PS feature mainly in the deliberation dynamics, which consists of the sequential random activation of connected clips in episodic memory. In that process, which is a biased random walk, the h -values function as unnormalised transition probabilities that may be grounded in physical connections of different strengths. On that view, the PS learning rule describes the growth of such connections.³⁰ The learning dynamics is composed of two different processes that, individually, can enhance or diminish the connections, corresponding to a loss term and gain term of a master equation as used in statistical physics, quantum optics, or chemistry. The loss term, proportional to a damping parameter γ , models attenuation of the h values towards a minimal, base-level connection. The gain term, on the other hand, models the growth of the h values in proportion to every positive reward R and glow g .

While the PS learning rule, as mentioned, does not refer to future states or actions, a notion of temporal non-locality enters implicitly through the glow term g from Eq. (5.6) in Eq. (5.18). This is particularly relevant for state-action pairs $(s, a) \neq (s^{(t)}, a^{(t)})$ that do *not* correspond to the transitions that were actually taken in round t . As we already mentioned, if the glow mechanism is active ($\eta < 1$), *all* connections that were used in a deliberation process remain susceptible for growth under future rewards, depending on the glow status of the corresponding edges in the ECM at the time when a reward is received. Glow describes a non-local memory effect different from the explicit non-locality that one has in Eq. (5.17). The effect of glow in PS is similar to the effect of eligibility traces in reinforcement learning (Sutton & Barto, 2018, Ch. 12).³¹ But the intuition behind the glow mechanism is more physical: glow can be understood as a kind of susceptibility that decays over time, which is found in many physical and biological systems.

Regarding the performance of PS as a learning model, we already mentioned that the learning rule in PS was not designed such as to converge to optimal behaviour, unlike standard reinforcement learning, where the value function through which

²⁹ See Chap. 4, Footnote 20, for pointers to philosophical literature critical of the assumption of a utility function. PS is, thus, not affected by this criticism.

³⁰ This is reminiscent of Hebbian learning in a neural network: connections of clips that are activated in sequence will grow, if the resulting action is rewarded. Different from neural networks is, however, the thoroughly stochastic dynamics in PS, which is modelled as the random walk of a single excitation through the clip network.

³¹ Both mechanisms can be seen as addressing the so-called credit assignment problem (Minsky, 1961).

optimality is defined is built into the formalism. Nevertheless, in many applications the learning performance of PS is competitive and comparable with standard reinforcement learning methods. Some of these applications will be described in more detail in Sects. 5.4 and 5.5. A detailed comparison of (two-layered, non-reflective) PS with the two reinforcement learning algorithms mentioned above, Q-learning and SARSA, can be found in Melnikov et al. (2018a) and Boyajian et al. (2020); see also Mautner et al. (2015); Bjerland (2015). The main findings of these studies can be summarised as follows. The basic version of PS, featuring simple two-layered clip networks, shows performance comparable with Q-learning or SARSA, at least in the standard textbook examples. These are simple computer games where an artificial agent is put in an idealised task environment such as *grid world*, *mountain car*, or the game *Infinite Mario Bros* (Bjerland, 2015). Typically, PS is fast in finding a fairly good solution to the task environment, while it does not always find the optimal solution; Q-learning and SARSA may take a little longer but typically find the optimal solution (Melnikov et al., 2018a). A study of the convergence properties of PS, together with a detailed formal comparison of PS with reinforcement learning, is given in Boyajian et al. (2020). From a user perspective, a relevant difference lies in the costs of optimising the model meta-parameters for specific task environments, for which PS requires a 1–2 orders of magnitudes smaller ensemble of agents (Melnikov et al., 2018a)), as well as the ease with which PS agents can adapt to non-stationary environments. While PS is perhaps not the most powerful reinforcement learning algorithm, it performs well on standard reinforcement learning problems, where it is flexible and easy to use.

For our context, however, the main difference between PS and standard reinforcement learning lies in the conceptual focus of the models, that is, in what they are meant to model. PS is not only about learning, rather, it is a model of memory and agency. Accordingly, unlike a Markov Decision Process, the PS agent’s clip network is not a description of a task environment *per se*, but more generally captures the agent’s experiences from its interactions with the environment. This may include a representation of the percept and action spaces, as in reinforcement learning, but it may also include other states of its memory that play a role in the deliberation dynamics, which PS resolves in detail.

While explicit goals and value functions are main features of reinforcement learning,³² there is no explicit goal in PS and thus no explicit value of a state (or action) in view of reaching that goal. Instead, the main features of PS lie in the model of episodic and compositional memory, its dynamics, and its integration into a model of learning and agency (Müller & Briegel, 2018). Given the form of the agent’s memory and its short and long-term dynamics, one focus of interest in PS

³² Sutton and Barto (2018, 3) highlight the existence of a goal as one of the key features of reinforcement learning: “All reinforcement learning agents have explicit goals, can sense aspects of their environments, and can choose actions to influence their environments.” The goal is defined, implicitly or explicitly, through the structure of the reward function. The latter, in turn, determines the *values* of different states or state-action pairs, i.e., how useful they are for the agent in order to reach the goal.

lies in the kinds of behaviour the agent can exhibit or develop over time. One type of behaviour that one observes is the basic capacity to learn. Another type of behaviour is a form of creativity that emerges under clip composition (Briegel, 2012). A third type of behaviour is the capacity of abstract concept formation (Eva et al., 2023), which will also be discussed in Sect. 5.5.

PS, as a model of agency, is thus different from reinforcement learning in many respects. It should be pointed out, however, that the theory of reinforcement learning, with its highly developed mathematical apparatus and methodology, can also provide valuable resources for the modelling of agency. For example, it may be possible to leverage PS with model-based methods from reinforcement learning, e.g., in order to integrate planning. On the other hand, reinforcement learning could benefit from the memory architecture of PS and its more advanced notions of deliberation and its focus on agency in open environments.

In the next section, we will discuss some applications of PS in quantum information science. These applications build on examples given in Chap. 3 and illustrate the way how artificial systems—most of which are still instances of a fairly thin notion of agency—are finding their way into modern quantum technologies. All applications we discuss use the implementation of the PS model introduced in Sect. 5.2, sometimes adapted to the specific applications, e.g., via concrete rules for clip composition and deletion or via the implementation and scaling of the rewards.

5.4 Applications of PS in Quantum Information Science

Modern applications of quantum physics encode information in the state of elementary quantum systems and use these systems for novel ways of information processing and communication. These applications require the precise control of (possibly complex) quantum systems and their internal dynamics. They offer a natural way to introduce a low-level notion of artificial agency, which is used when people talk, e.g., about different parties in a communication protocol or a controller in quantum error correction. Of course, this usage does not yet include the full scale of explorative experimenting, modelling, and scientific discovery. But it includes adaptive control and reinforcement learning.

Many quantum information applications use highly entangled states, such as the Bell singlet state that we already met in Sect. 3.4.3 in the context of Bell's theorem and the generation of device-independent randomness. In this section, we will consider further examples of applications in the field of quantum information in which a minimal notion of agency is used. We discuss applications in which a PS agent is used to control and manipulate the entangled quantum states used in quantum error correction, quantum communication, and measurement-based quantum computation, as well as applications to the design of photonic experiments. The technical background for our discussion in what follows is the description of Bell-type states as a quantum information resource in Sect. 3.4.4 and the description of the quantum error correction protocol in Sect. 3.5.4. Readers unfamiliar with that

background can still obtain a general impression of simple instances of PS at work by skimming the following material.

5.4.1 *Quantum Error Correction: In-situ Optimisation of Quantum Codes in Noisy Environments*

A quantum computer is a physical system that is used to process information encoded in the state of elementary quantum systems, e.g., in the form of qubits, as described in Sects. 3.3.3.2 and 3.4.4. For practical applications, this will require the precise control of the state of a large number of qubits, which constitute some kind of quantum memory. In the standard model of quantum computation one needs to precisely control the quantum memory's time evolution, generated by physical interactions of the qubits with each other and with external electromagnetic control fields. This represents a major challenge since, at the same time, the quantum memory needs to be shielded from unwanted interactions of the qubits with the environment. Since no physical system can be perfectly isolated, one will always have to cope with some form of (quantum) noise that affects the state of the physical qubits that constitute the quantum memory, as described in Sect. 3.5.4.1.

To stabilise the temporal evolution of the quantum memory against the detrimental effects of noise, one can apply methods of quantum error correction as described in Sect. 3.5.4. The basic idea is to encode every logical qubit into (a two-dimensional subspace of the Hilbert space of) a larger number of physical qubits, thereby delocalising the information. The study of suitable quantum error correcting codes for different physical implementations is an active field of research of its own.³³ Typically, these codes are designed and optimised with respect to a specific quantum device and its environment while also taking into account hardware restrictions, such as the number of qubits that can be controlled coherently simultaneously, reflecting the technological state-of-the-art.³⁴ Quantum error correction is then a fully scripted protocol based on a fixed quantum code which can be executed via a fully automatised scheme.

In the application to be discussed here, we will abandon this assumption and consider a scenario where the model of the environment and of the hardware is up for revision. We will consider a PS agent that does not assume a fixed model but instead learns how to adapt and change a given code to the conditions that prevail in a given device. The PS agent can model a human experimenter who reacts to a change of the experimental conditions that lead to an increasing error rate; or it can describe an artificial agent that is designed to assist to stabilise the fault-tolerant operation of a quantum computer.

³³ For recent results, see, e.g., Grassl (2023) and Bravyi et al. (2024).

³⁴ A generic characterisation and summary of the current technological situation goes under the term *Noisy Intermediate Scale Quantum* (NISQ) era (Preskill, 2018).

This application can be considered to be a first step towards *model building*, as discussed in Sect. 4.4, of an agent that interacts with its environment. In this case, the model of the environment is implicit in the error correcting code that matches the targeted low error rate, and thus in the effect of the agent's actions (code deformations) on the environment and the system it controls. For a proper discussion of this application we first need to briefly consider a special case of stabilizer codes (Sect. 3.5.4.3) known as surface codes.

Surface Codes A class of stabilizer codes that are considered among the leading candidates for the realization of fault-tolerant quantum computers are the topological *surface codes* (Dennis et al., 2002; Fowler et al., 2012a), with the toric code and the planar codes being two prominent examples. To visualise the surface code, one can imagine the physical qubits to be arranged on the edges of a lattice embedded in a two-dimensional surface as depicted in Fig. 5.7. The surface can be a plane or a torus, whence the names of the different codes. Furthermore, the lattice does not need to be regular, but can be replaced by any tessellation of the surface. Importantly, the tessellation significantly affects the code performance with respect to the error model, which in turn depends on the device.

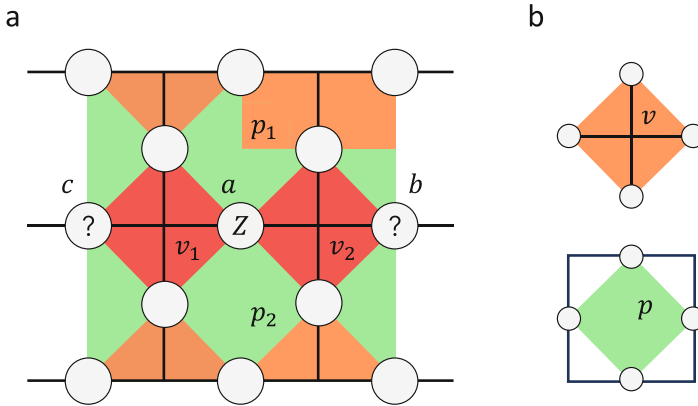


Fig. 5.7 Surface code on a square lattice. **(a)** Physical qubits, indicated as grey dots, are located at the edges of a square lattice. **(b)** With every vertex v (at the centers of the red and the orange squares) of the lattice one associates stabilizer observables M_v and M_p (Eq. 5.20), respectively. M_v acts on qubits located at the four edges incident on vertex v , while M_p acts on the qubits located at the four edges forming the boundary of plaquette p . **(a)** A single phase flip error Z_a on qubit a (grey dot between the red squares at the center of the lattice) transforms the logical qubit into $Z_a|\psi\rangle_L$, which changes the eigenvalues of stabilizer observables M_{v_1} and M_{v_2} whose support includes qubit a . Similarly, a single bit flip error X_a on qubit a changes the eigenvalues of stabilizer observables M_{p_1} and M_{p_2} . When correcting for an error, say Z_a , one only has the information about the error syndrome, i.e., the changed eigenvalues of M_{v_1} and M_{v_2} , available. However, an error $Z_c Z_b$ would have resulted in the same syndrome. Under the assumption that a single error is more likely to occur than two, one corrects for an error Z_a rather than $Z_c Z_b$.

Consider the situation displayed in Fig. 5.7a where qubits (indicated as grey dots) are located on the edges of a regular lattice (which is a special instance of a tessellation). The stabilizer observables (generators) of the surface code have a simple geometrical description associated with the lattice. There are two types of observables M_k , one is associated with the vertices of the lattice ($k = v$) and the other with its faces, or *plaquettes* ($k = p$):

$$\begin{aligned} M_v &= \prod_{j \in N_v} X_j, \\ M_p &= \prod_{j \in N_p} Z_j, \end{aligned} \quad (5.20)$$

where N_v denotes the set of all qubits associated with the edges adjacent to vertex v , and N_p all qubits associated with the edges forming the boundary of plaquette p (Fig. 5.7b). Note that the index j in Eq. (5.20) labels individual qubits, while the indices v and p are purely geometrical labels that indicate which qubits are grouped together in the corresponding stabilizer operator. One can easily check that all M_k pairwise commute (which is, in fact, true for any tessellation). When the agent measures all eigenvalues m_k , constituting the error syndrome, there are two possible outcomes. Either all of the eigenvalues are $+1$, in which case there is no correctable or detectable error and the encoded qubit is still in the code space.³⁵ Or at least one of the eigenvalues is -1 , in which case the encoded qubit is rotated into a subspace orthogonal to the code space. If the eigenvalue of one of the stabilizers is found to be -1 , this indicates that an error occurred on one of the physical qubits in the support of that stabilizer. For example, suppose that a single phase flip error Z_a occurred on qubit a , which transforms the logical qubit into $Z_a|\psi\rangle_L$. Qubit a belongs to the neighbourhood of two vertices v_1, v_2 and two plaquettes p_1, p_2 (see Fig. 5.7a). For the corresponding vertex and plaquette operator we then have

$$\begin{aligned} M_{v_1} Z_a |\psi\rangle_L &= -Z_a |\psi\rangle_L, \\ M_{v_2} Z_a |\psi\rangle_L &= -Z_a |\psi\rangle_L, \\ M_{p_1} Z_a |\psi\rangle_L &= +Z_a |\psi\rangle_L, \\ M_{p_2} Z_a |\psi\rangle_L &= +Z_a |\psi\rangle_L. \end{aligned} \quad (5.21)$$

That is, the appearance of negative eigenvalues of the stabilizers M_{v_1}, M_{v_2} but not of M_{p_1}, M_{p_2} , reveals that at least one Z -error has happened. Similarly, an error X_a would be indicated by the eigenvalues of M_{p_1}, M_{p_2} changing sign, but not of

³⁵ An alternative way of thinking about it is to say that the reduced (mixed) state of the memory is projected back into the code space or one of the orthogonal error spaces by measuring stabilizers. In that way, unwanted coherent evolutions—sometimes also referred to as coherent errors—can be interpreted as Pauli channels.

M_{v_1}, M_{v_2} while Y_a is indicated by the eigenvalues of all four stabilizers changing sign.

In this way, the measured eigenvalues of the stabilizers do not only signal which type of error has happened, but one also has a likelihood estimate of where the error happened. In Fig. 5.7a one can see two possible ways how a given error syndrome could have occurred, namely either through an error on qubit a , or two errors on qubits b, c . If one assumes that a single error is more likely to occur than two errors, one corrects for an error on qubit a . This is an instance of decoding and correction based on the so-called minimum-weight perfect matching algorithm (Fowler et al., 2012b; Edmonds, 1965).

Overall, to implement quantum error correction, the agent needs to first measure the eigenvalues of all generators M_k of the stabilizer code (see Sect. 3.5.4). From the set of eigenvalues m_k obtained, i.e., from the error syndrome, the agent can then find out where an error most likely happened, of which type it was, and what it needs to do to correct it. Obtaining this information from the error syndrome will require some classical computation. We will not discuss the decoding procedure in detail, but beyond the minimum-weight perfect matching algorithm just mentioned, there exist more efficient algorithms such as the union find decoder (Delfosse & Nickerson, 2021). Note that failure to pick the right correction can lead to a logical error and, consequently, to the failure of the memory. To increase the likelihood of making the right choice, one can however increase the size of the lattice.

With this procedure, quantum information can thus be stored and protected against noise acting on individual qubits. This requires repeated rounds of error correction, including syndrome measurement, decoding, and unitary correction operations. The lifetime of the logical qubit is thereby increased, similar as in the context of the repetition code (Eq. (3.163)). To keep the number of required rounds of error corrections small, it is desirable to use a surface code with a small logical error probability. For the surface code on a regular lattice, one way of achieving this is to increase the size of the surface, similar as for the repetition code, where one can decrease the logical error rate by increasing the number of physical qubits. However there are other, more sophisticated ways that we describe next.

The quantum repetition code and the surface code are two examples of quantum codes that are highly symmetric, which helps to analyse their behaviour under simple noise models, but they may not be the best choice in a real environment. For a concrete physical environment, the types of errors and their corresponding rates may depend on the spatial location of a qubit in the memory, and errors on different physical qubits may be correlated. Some of the error sources may be unknown, or they may change over time. Even if the error sources and characteristics are known, it may be difficult to find the best error correcting code in terms of resource requirements and performance given a specific hardware. A method to deal with such a more general situation is to consider the larger space of surface codes that can be obtained from the initial symmetric code by a set of basic *code deformations* that change the tessellation of the surface and thus the performance of the code. Mathematically, these codes are generated by two types of transformations that change the local tessellation of the surface, where one type is associated with every

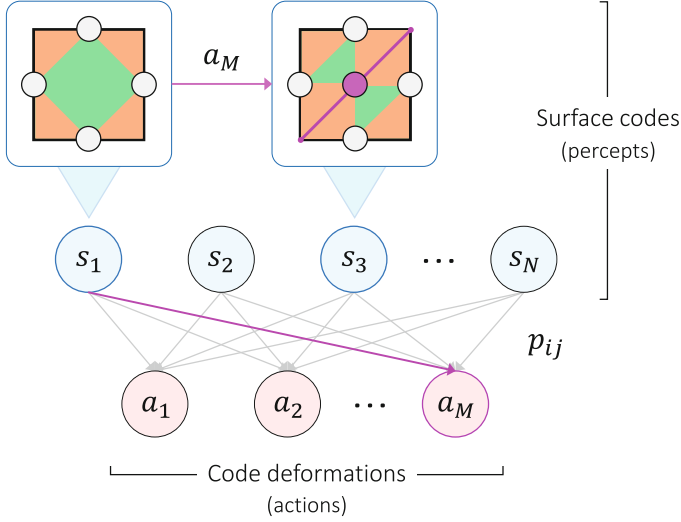


Fig. 5.8 Clip network of a PS agent acting on quantum surface codes. Percept clips are coloured *blue* and represent surface codes. Action clips are coloured *red* and represent code deformations. If action a_M is performed upon encountering percept s_1 (indicated by a *pink arrow*), the code associated with s_1 is deformed according to a_M . This results in a new code, for example s_3 , that is encountered by the agent in the next step. The insets show how a single plaquette from the surface (code) changes under the deformation. Note that the set of available actions (i.e., code deformations) depends on the tessellation (i.e., the code) encountered. In the ECM this means that not every percept is necessarily connected to every action

vertex and the other with every plaquette (Dennis et al., 2002). An example of a code deformation is shown in Fig. 5.8, where a plaquette of the lattice is split into two faces by adding an edge, together with a new qubit. Physically, code deformations are thus realised by adding physical qubits to the surface and entangling them with the other qubits in a specific way.³⁶ Note that these are similar physical operations with which the agent can generate the initial surface code, in the first place.³⁷ Using these operations, the agent can search for a code that has an acceptable logical error rate for a simplified noise model that is thought to be close to the real environment and then further adapt the code to errors generated by the concrete physical environment or device. There may also be additional resource constraints, such as a maximum available number of physical qubits or their interaction range.

³⁶ Similarly, code deformations can also be achieved by merging faces of a tessellation (and thus removing physical qubits).

³⁷ Theoretical proposals and proof-of-principle experiments which demonstrate the realization of surface codes exist for different physical platforms, including neutral atoms in an optical lattice (Joo et al., 2013), trapped ions (Erhard et al., 2021; Nigg et al., 2014), or superconducting qubits (Krinner et al., 2022).

To find a code that satisfies all given constraints is generally not a simple task but it can be tackled with ideas from reinforcement learning, specifically with Projective Simulation, as was shown by Poulsen Nautrup et al. (2019). The main idea is to consider the environment of the agent as a discrete space, where each point (or state) of that space corresponds to a surface code. The actions of the agent then consist of basic moves applied to any of the faces of a tessellation. These actions constitute the code deformations illustrated in Fig. 5.8. They connect different surface codes, i.e., states of the environment. The agent is rewarded if it finds a code with a logical error rate below a specified target value. Stated in this way, the learning task is similar to *grid world* as described in Sect. 5.2.4. The clip network of a PS agent operating in the surface-code environment is shown in Fig. 5.8.

It is interesting to note that the set of quantum codes can be viewed both as a purely mathematical set, in which case the actions correspond to mathematical operations, and as a set of physical states of matter constituting a quantum memory, all encoding the same logical information. In the latter case, the actions are physical operations applied by the agent in the laboratory, for example, using laser pulses in an ion-trap implementation. As already mentioned, this dual perspective can be used by an artificial agent for off-line simulations of an experiment, using a model of the environment described by a certain noise channel; but it can also be used for online control and optimisation in a given, real environment whose precise noise characteristic is unknown. Since the PS agent exhibits some form of transfer learning (Poulsen Nautrup et al., 2019), a combination of the two modes of usage provides a means of accelerating the code optimisation for specific hardware realisations and constraints.

5.4.2 *Quantum Communication Through Noisy Channels: Discovery and Variation of Quantum Protocols*

In quantum communication, one goal is to transmit quantum information between two different locations connected by a physical channel. Typically one considers two parties, Alice (A) and Bob (B), located at different places, who communicate by exchanging quantum systems in specific states. Their main problem is that due to unwanted interactions of the carrier of the quantum information with degrees of freedom of the channel, noise generally disturbs the transmission, i.e., the quantum channel is imperfect.

To deal with the effect of noise, one could in principle use methods of quantum error correction as described above, so that quantum communication would simply be another application: one would encode the logical information in the quantum state of several physical qubits, which would then be sent through the channel. Formally, these two applications of quantum error correction are very similar. In the case of quantum memory discussed in Sect. 5.4.1, the channel is realised by uncontrolled interactions of a spatially localised qubit, such as an atom in a trap,

with its environment, constituted, e.g., by fluctuating electromagnetic forces that act on the atom. In the quantum communication case here at issue, we envision a qubit that moves through space, such as a photon propagating through an optical fibre or through the atmosphere. The environment is then represented by the fibre and its specific properties (refractive index, temperature) that affect the photon wave packet on its way between Alice and Bob.

The modelling of the environment in these two cases is closely related. For example, suppose that Alice and Bob exchange a two-state quantum system, i.e., a quantum bit, and the environment is a depolarising (Pauli) channel as described by Eq. (3.156) with $p_1 = p_2 = p_3 = p/4$. Assuming a Markovian environment, the noise parameter p would typically have an exponential dependence in both cases, either on time $p = p(t) = 1 - \exp(-t/t_0)$ (quantum memory), or on distance, $p = p(l) = 1 - \exp(-l/l_0)$ (quantum communication), where t and l measure storage time and communication distance in units of t_0 and l_0 , respectively. If Alice prepares a qubit a' in the state $\rho_{a'} = |\psi\rangle_{a'}\langle\psi|$ and sends it through the channel, Bob will receive a mixed state of the form

$$\rho_{a'}(p) = (1 - p)|\psi\rangle_{a'}\langle\psi| + \frac{p}{2}I_{a'}. \quad (5.22)$$

One way of addressing the communication problem would be to encode the quantum bit into a larger group of photons, send this group through the channel, and apply quantum error correction at designated repeater stations along the way. However, there is a more elegant approach that uses quantum teleportation (Bennett et al., 1993) as described in Sect. 3.4.4: Suppose that Alice and Bob share an entangled state $|\Phi^+\rangle_{ab}$ of two physical qubits, one held by Alice (qubit a) and the other by Bob (qubit b). In addition, Alice has another qubit (a') whose quantum state $|\psi\rangle_{a'} = \alpha|0\rangle_{a'} + \beta|1\rangle_{a'}$ she wants to convey to Bob. As discussed in Sect. 3.4.4, instead of sending qubit a' physically to Bob (and therefore through the noisy channel), Alice can teleport the quantum state $|\psi\rangle$ to Bob by means of a Bell measurement that she applies to her qubits a and a' , and by communicating the result of this measurement (two bits of classical information) to Bob, who can then use this information to transform his qubit b into the state $|\psi\rangle_b = \alpha|0\rangle_b + \beta|1\rangle_b$. The same logical information (represented by the amplitudes α and β) that was initially carried by Alice's qubit a' is now carried by Bob's qubit b .

Having a shared Bell pair as a resource, Alice and Bob can thus circumvent the noise channel. Instead of sending the quantum information through the channel, they teleport it and use the quantum correlations, i.e., their shared non-local Hilbert space, as an alternative communication channel. One obvious problem is: how do Alice and Bob get a Bell pair, if they are not endowed with this resource from the beginning? Alice could, of course, entangle two qubits (a and b) in her laboratory to generate the Bell state $\rho_{ab} = |\Phi^+\rangle_{ab}\langle\Phi^+|$, and then send one of the qubits (b) to Bob. While the qubit b travels to Bob, it will be exposed to the noisy channel, which degrades the entanglement between a and b . For the example of the depolarizing

(Pauli) channel, Eq. (3.156), this will result in a mixed two-qubit state of the form³⁸

$$\rho_{ab}(p) = (1 - p)|\Phi^+\rangle_{ab}\langle\Phi^+| + \frac{p}{4}I_a \otimes I_b. \quad (5.23)$$

As long as the noise is not too large, i.e., for $0 \leq p < 2/3$, the state ρ_{ab} of Eq. (5.23) still contains some entanglement. However, there is no advantage if Alice and Bob use this mixed entangled state directly for teleportation. If they perform the teleportation protocol with ρ_{ab} of Eq. (5.23), Bob's qubit will be transformed into the state

$$\rho_b(p) = (1 - p)|\psi\rangle_b\langle\psi| + \frac{p}{2}I_b, \quad (5.24)$$

which is identical with Eq. (5.22), which means that the effect is the same as if Alice had simply sent her qubit a' through the channel to Bob.

But there is a solution. Another key protocol of quantum information—next to quantum teleportation—is *entanglement purification* (Bennett et al., 1996).³⁹ If Alice and Bob share *two* identically prepared states of the form $\rho_{ab} = \rho_{ab}(p)$ with $p < 2/3$, they can prepare, with probability larger than $1/4$, one state of the same form $\rho_{ab}(p')$ but with higher purity (less noise) $p' < p$.⁴⁰ To achieve this, they need only local operations in their respective laboratories and classical communication. If Alice and Bob thus share a sufficiently large ensemble of such states, by iterating this stochastic procedure on properly selected subensembles of states with increasing purity they can distill, from the initial ensemble, states with almost perfect purity, i.e., with $p \rightarrow 0$ and $\rho_{ab} \rightarrow |\Phi^+\rangle_{ab}\langle\Phi^+|$.⁴¹

Quantum teleportation and entanglement purification are two key protocols of quantum communication.⁴² In the quantum repeater proposed by Briegel et al. (1998) both protocols are generalized and combined into a nested scheme, which solves the problem of long-distance quantum communication in the general case when all physical processes involved—including the local operations applied by the agents—are noisy. It has been shown that the number of uses of the channel

³⁸ If we replace $|\Phi^+\rangle_{ab}$ with the singlet state $|\Psi^-\rangle_{ab}$ from Eq. (5.23), the resulting state is also called a Werner state (Werner, 1989).

³⁹ See also Deutsch et al. (1996).

⁴⁰ For an introduction into the principles of entanglement purification see, e.g., Briegel (2001).

⁴¹ The protocol is remarkable beyond its immediate applications in quantum communication insofar as states $\rho_{ab}(p)$ as in Eq. (5.23) admit a local hidden variable model for $1 - p \leq 1/\sqrt{2} \simeq 0.71$ (Werner, 1989) while they are already distillable for $1 - p > 1/3$. Thus, if Alice and Bob have an ensemble of mixed states with $1/3 < 1 - p \leq 1/\sqrt{2}$, each of which admits a hidden variable model, they can distill out of this ensemble a smaller ensemble of states with $1 - p > 1/\sqrt{2}$ none of which admits such a model (see also Popescu 1994).

⁴² These protocols can also be understood as a measurement-based computation with ensembles of graph states, providing further insights regarding error thresholds and distillation yields in the practically important case of *noisy local operations* (Zwenger et al., 2016).

segments between the repeater stations grows only polynomially with the total distance between Alice and Bob. In contrast, in a naïve approach based on direct transmission this number would grow exponentially with the distance.

As was shown by Wallnöfer et al. (2020), a PS agent can learn the basic protocols for quantum communication, including teleportation and entanglement purification, and their combination into the quantum repeater. The specification of the task environment in each of these cases includes a set of available actions and possibly other resources. For example, for teleportation, the task is to transfer an arbitrary quantum state $|\psi\rangle_{a'}$ of a qubit a' held by Alice to Bob, without sending the physical qubit a' itself through the channel. Depending on the complexity of the challenge, Alice and/or Bob are equipped with further resources. One possibility is to equip them with an entangled Bell state $|\Phi^+\rangle_{ab}$ of two further qubits a and b , held by Alice and Bob, respectively. Another possibility is to equip only Alice with one or two auxiliary qubits, which she is allowed to send to Bob. Beyond these resources, Alice and Bob can act locally on their qubits through a basic set of operations, which includes certain (one- and two-qubit) unitary operations, together with the capacity to measure any qubit in the computational basis, corresponding to measurements of the observables $Z_{a'}$, Z_a , Z_b . Since the measurements may be applied to entangled states, the outcomes of some of these actions will be *indeterministic* (Sect. 3.4.3). The agent is thus situated in a stochastic environment and has to learn the appropriate next action for all branches of the different measurement outcomes. As was shown by Wallnöfer et al. (2020), for all of the mentioned scenarios, the PS learning framework can be used to solve the task with the given resources. In the case where Alice and Bob are provided with a shared entangled Bell state as a resource, the PS agent learns to teleport the state with a minimal number of actions that is comparable to the original protocol of Bennett et al. (1993).

To deal with larger-scale quantum channels, Wallnöfer et al. (2020) consider a more complex scenario where the agent can use teleportation and purification as composite actions and delegate sub-tasks to other agents of the same type, similar to the idea of clip composition and hierarchical skill learning (Hangl et al., 2016), respectively. This allows for the possibility of nesting, as it is used in the quantum repeater protocol (Briegel et al., 1998). The capacity to delegate is particularly valuable in inhomogeneous scenarios, where the noise that acts on different parts of the channel is different. In this case, the PS agent was shown to outperform the original repeater protocol that assumed homogeneous, or symmetric, noise channels. This situation is thus similar to the PS-controlled adaptation of the surface code in a quantum memory subjected to noise, where the noise may depend on time and on the location of the physical qubits in the code.

In the specification of the available resources and actions, Wallnöfer et al. (2020) admittedly make use of existing knowledge. One can argue (as these authors in fact do) that the essential step in the discovery of quantum teleportation was to *conceive of the very idea* that entanglement and the puzzling quantum correlations described by Bell's theorems (see Sect. 3.4.3), which previously had only been addressed in foundational and philosophical discussions on the interpretation of

quantum mechanics, could somehow be useful for a new kind of communication. Only agents like humans, who can draw from a large research context, seem capable of forming such an idea. But even if that is the case, the remaining essential task of working out the details can still be difficult. And these difficulties will definitely increase once the problems become more complex, as for example in concrete physical implementations of large-scale quantum networks.

5.4.3 *Measurement-based Quantum Computation: Implicit Model-building of the Dynamics of a Single Qubit*

In this section, we focus on a situation where the objective can be summarised as experimenting with a quantum system to learn its dynamical behaviour. The task of the agent is to make a correct prediction about the phenomenon that it will observe when it interacts with the system for the next time. If the prediction turns out to be correct, the agent is rewarded. This objective can be considered as a basic element of human experimenting, where the first step is to familiarize oneself with the behaviour of the system, learning how it reacts to our interventions. After all, one criterion for *understanding* a system is that the agent consistently confirms her expectations about the system's behaviour.

The system we consider here is a single qubit, or spin-1/2 system, that is subject to a weak magnetic field $\vec{B}$ of an otherwise unknown magnitude and direction. The Hamiltonian of that system (Eq. (3.81)) is given by $H = -\gamma \vec{B} \cdot \vec{S}$, as discussed in the context of the Stern-Gerlach experiment in Sect. 3.3.3.3. The knowledge of the intrinsic dynamics of the system, which may be caused by residual magnetic (stray) fields in the agent's experiment, is relevant for most protocols of quantum information including the ones we already discussed. It is also important for measurement-based quantum computation (MBQC), where single-qubit measurements on an entangled cluster state drive the entire quantum computation (Raussendorf & Briegel, 2001). That is, any quantum algorithm is implemented by an adaptive, spatio-temporal pattern of measurements. For this to work, it is important that the changes on the system and its environment are only due to the measurements and not due to processes that are not controlled by the agent. If there are uncontrolled interactions of the system with other degrees of freedom of the environment, one can still apply methods of quantum error correction, like the ones we discussed in the previous Sections. However, if these interactions are due to static (or slowly changing) fields, as is the case with magnetic stray fields, there is a more economic way of dealing with the situation.

In the Heisenberg picture, the static fields lead to a rotation of the spin (vector) observable $\vec{S}(t) = U^\dagger(t) \vec{S}(0) U(t)$ of every qubit, as discussed in Sect. 3.3.3.3. The unitary $U(t) = \exp(-\frac{i}{\hbar} H t)$, which is generated by the Hamilton operator as in Eq. (3.81), depends on the strength and direction of the field $\vec{B}$ at the location of each qubit. The eigenstates of the spin observables change accordingly, including

the eigenstates of $S_z = \frac{\hbar}{2}\sigma_z$, which define the computational basis of the qubit as in Eq. (3.77).

A one-way quantum computation is driven by a series of measurements (in different directions) of the qubits of a fixed cluster state. If the basis of a qubit changes over time, from one measurement to the next, the agent can still perform the quantum computation and execute the desired quantum algorithm—if it adapts the measurement direction to the basis change. But for this it has to *know* the underlying dynamics of the qubit, or the strength of the $\vec{B}$ field that acts on that qubit.⁴³

To acquire knowledge of the stray fields, the agent could be endowed with some extra device or procedure that it could use to map out the magnetic fields in its experiment. In fact there are various methods that could be used for this task, known, e.g., as parameter estimation, Hamiltonian estimation, or quantum sensing.⁴⁴ In the following, we will focus on the simplest scenario where the agent operates on a single qubit exposed to some unknown magnetic field. Connecting to the discussion in previous Chapters, we are specifically interested in how the agent can (implicitly) build a model of the system while interacting with it. This scenario is thus meant to illustrate the process of *experimenting*. The goal of the PS agent is not to provide an explicit estimate of the magnetic field and it is not optimized towards high-precision applications. We do not provide additional resources to the PS agent, except the following basic capacities.

- **Actions.** The agent can measure the qubit in different directions $\vec{n}$, corresponding to the observables $\vec{n} \cdot \vec{\sigma}$. In the following, we identify the actions of the agent with the observables that it measures.
- **Percepts.** A percept is the specification of the measured eigenvalue ± 1 together with the chosen observable. In the scenario on which we focus in what follows, where each interaction round starts with an identically prepared qubit, the percept can be replaced by a simple trigger signal.
- **Rewards.** The agent is rewarded if it finds eigenvalue $+1$.

To simplify the analysis, we assume that the magnetic field points in the z -direction, i.e., $\vec{B} = B\vec{e}_z$. We also assume that the agent knows the direction but not the

⁴³ An expensive way of dealing with this situation, that is, to perform the quantum computation without knowledge of the stray fields, would be to treat the environment (the stray fields) as a noisy channel and use quantum error correction. This amounts to performing a fault-tolerant one-way quantum computation. What we are considering here is more economic, at least if the effect of the stray fields, in terms of the so-called coherent error rate, is comparable to or larger than the incoherent error rate due to noisy interaction. In such a situation, the current scenario offers a way to deal with the unitary part of the environmental dynamics (which is assumed to be quasi static compared to the time scale of the fluctuating dynamics that creates incoherent errors) in an adaptive, resource efficient way, while the quantum error correction is reserved for the incoherent errors.

⁴⁴ All of these protocols, one could argue, can be considered as instantiations of the modelling loops described in Sect. 4.4.1. The precise detection and estimation of weak magnetic fields has many potential applications outside quantum information processing, e.g., in chemistry and biomedicine (Aslam et al., 2023; Briegel et al., 2025).

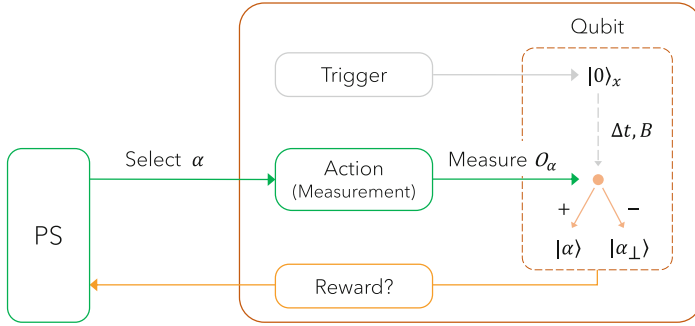


Fig. 5.9 Set-up of a PS agent that learns how to adapt its measurements of a qubit whose initial state $|0\rangle_x$ changes due to the presence of some unknown magnetic field acting on the qubit. Each round of the experiment is triggered by the preparation of the qubit in the state $|0\rangle_x$. After a fixed time interval Δt , the qubit is measured along one of several different directions $\vec{n}_\alpha = \cos \alpha \vec{e}_x + \sin \alpha \vec{e}_y$, corresponding to different observables $O_\alpha = |\alpha\rangle\langle\alpha| - |\alpha_\perp\rangle\langle\alpha_\perp|$, as in Eq. (5.28), and $\alpha_\perp = \alpha + \pi$. A reward is issued if the measured eigenvalue is +1 and the ECM of the PS agent is updated accordingly. Adapted from Tiersch et al. (2015)

magnitude B of the magnetic field, which we write in the form $B = \omega/|\gamma|$, using the (Larmor) frequency ω as in Eq. (3.83).⁴⁵ We consider the following scenario describing a series of identical experiments (see Fig. 5.9). In each experiment, the agent first prepares, at time $t = 0$, the qubit in an eigenstate $|0\rangle_x = (|0\rangle + |1\rangle)/\sqrt{2}$ of $X(0) = \sigma_x(0)$. The density operator of the qubit thus has the form $\rho = |0\rangle_x\langle 0| = (I + X(0))/2$. Then, after some fixed time interval Δt during which the qubit is exposed to the magnetic field, the probability for obtaining +1 upon measuring the spin in the x -direction will generally no longer be one:

$$\Pr(\uparrow, x; \Delta t) = \text{tr}(\rho \frac{I + X(\Delta t)}{2}) = \text{tr}(\frac{I + X(0)}{2} \frac{I + X(\Delta t)}{2}) < 1. \quad (5.25)$$

The agent now measures the qubit's spin in some direction $\vec{n}$, which corresponds to the observable $\vec{n} \cdot \vec{\sigma}(\Delta t)$. The probability to obtain the eigenvalue +1 is given by

$$\Pr(\uparrow, \vec{n}; \Delta t) = \text{tr}(\frac{I + X(0)}{2} \frac{I + \vec{n} \cdot \vec{\sigma}(\Delta t)}{2}). \quad (5.26)$$

The question is which direction $\vec{n}$ the agent needs to choose to obtain the result +1, and thus to be rewarded with certainty. For the specific set-up described, the direction will be somewhere in the xy -plane of the Bloch sphere, i.e., $\vec{n}_\alpha = \cos \alpha \vec{e}_x + \sin \alpha \vec{e}_y$, for some value of α . The action of the agent thus corresponds to

⁴⁵ These assumptions can be relaxed and the scenario can be generalised to the situation when both the direction and the strength of the magnetic field are unknown to the agent (Glesener, 2020).

the measurement of an observable

$$O_\alpha = \vec{n}_\alpha \cdot \vec{\sigma}(\Delta t) = \cos \alpha X(\Delta t) + \sin \alpha Y(\Delta t). \quad (5.27)$$

Using the spectral decomposition of this observable, it can be written in the form

$$O_\alpha = |\alpha\rangle\langle\alpha| - |\alpha_\perp\rangle\langle\alpha_\perp|, \quad (5.28)$$

with

$$|\alpha\rangle = \frac{1}{\sqrt{2}} \left(|0; \Delta t\rangle + e^{i\alpha} |1; \Delta t\rangle \right) \quad \text{and} \quad |\alpha_\perp\rangle = |\alpha + \pi\rangle \quad (5.29)$$

being the two eigenstates of O_α at the time Δt . (The states $|0; \Delta t\rangle$ and $|1; \Delta t\rangle$ are the computational basis states of the qubit at time Δt , i.e., the eigenstates of $Z(\Delta t)$). Similarly, one can write the observable $X(0)$ that appears in Eq. (5.26) in the form

$$X(0) = U(\Delta t)X(\Delta t)U^\dagger(\Delta t) = |\varphi\rangle\langle\varphi| - |\varphi_\perp\rangle\langle\varphi_\perp|, \quad (5.30)$$

with $|\varphi\rangle = U(\Delta t)|0; \Delta t\rangle_x$. Up to a global phase factor this state can also be written as

$$|\varphi\rangle = \frac{1}{\sqrt{2}} \left(|0; \Delta t\rangle + e^{i\varphi} |1; \Delta t\rangle \right) \quad \text{and} \quad |\varphi_\perp\rangle = |\varphi + \pi\rangle, \quad (5.31)$$

with $\varphi = \omega\Delta t$. If the expressions obtained for O_α and for $X(0)$ are inserted into Eq. (5.26), one can easily check that

$$\text{Pr}(\uparrow, \vec{n}_\alpha; \Delta t) = |\langle\alpha|\varphi\rangle|^2, \quad (5.32)$$

which will be maximised when $\alpha = \varphi$.⁴⁶ Choosing the right action can thus be rephrased as guessing a value of α that is closest to the phase φ in the state of Eq. (5.31).

We assume that the agent has a finite set of actions $a \in A$ available. Again to simplify the analysis, we assume that the set consists of four possible actions, corresponding to the measurement of the observables $X, Y, -X, -Y$ indicating the alignment of the spin vector along the spatial directions $\vec{e}_x, \vec{e}_y, -\vec{e}_x, -\vec{e}_y$. In the notation of Eq. (5.27), these choices correspond to O_α with $\alpha \in \{0, \pi/2, \pi, 3\pi/2\}$,⁴⁷ with two possible eigenvalues ± 1 in each case.

⁴⁶ The derivation of this result via the spectral decompositions of O_α and $X(0)$ can be obtained more straightforwardly by expressing the density operator, in the Heisenberg picture, as a function of the spin observables at time Δt . The derivation here is meant as a service to the reader to connect the treatment reported in Tiersch et al. (2015), which is based on the Schrödinger picture—as are most treatments in quantum information—to the picture that we use in this book.

⁴⁷ In other words, $O_{\alpha=0} = X, O_{\alpha=\pi/2} = Y, O_{\alpha=\pi} = -X, O_{\alpha=3\pi/2} = -Y$.

In these experiments, a reward is given after every measurement: $R = 1$ if eigenvalue $+1$ is obtained, $R = 0$ if the eigenvalue is -1 . Once the reward is issued, the h -values of the PS agent's memory are updated and the experiment is re-initialized with the same initial state. That is, the agent prepares the system again in state $|0\rangle_x$. Formally, the preparation can be described as a process where the agent measures the observable X of the qubit and finds the eigenvalue $+1$. At this point, the clock starts and the agent lets the system evolve for the time Δt after which it measures it again along one of the four different directions. The choice of the observable can be interpreted as a guess of the agent in which direction the spin vector will point after time Δt . If the agent finds the eigenvalue $+1$ again, it is rewarded, which can be interpreted as a satisfying confirmation that the guess was correct. If that is the case, the corresponding edge in the clip network will be strengthened, that is, the h -value increases accordingly.

The ECM of the PS agent dealing with this situation has four actions O_α as described above and a single percept, $s = \star$, a trigger indicating to the agent that the qubit has been initialized in state $|0\rangle_x$ —which is the eigenstate corresponding to eigenvalue $+1$ of observable $X(0)$. The weight of an edge connecting percept clip $\star$ with the action clip for O_α is here denoted by $h_\alpha = h_{\alpha,\star}$ with a single subscript. As the agent repeats the experiment, the h -values are updated according to the PS learning rule (Eq. (5.19)), which in this case reads

$$h_\alpha^{(t+1)} = h_\alpha^{(t)} - \gamma (h_\alpha^{(t)} - 1) + R^{(t)}, \quad (5.33)$$

where $R^{(t)} = 1$ for those values of α for which the corresponding observable O_α was actually measured in round t and a positive eigenvalue $+1$ was found. In all other cases, $R^{(t)} = 0$.⁴⁸ The probability of choosing action O_α in round t is given by the standard probability rule of the PS model,

$$p^{(t)}(\alpha) = \frac{h_\alpha^{(t)}}{\sum_{\alpha'} h_{\alpha'}^{(t)}}. \quad (5.34)$$

For a discrete set of angles $\{\alpha | O_\alpha \in A\}$, corresponding to a set of measurements (actions A) that the agent can perform, typically none of these actions will lead to a reward with certainty, because the corresponding $+1$ eigenstates $|\alpha\rangle$ will normally not perfectly coincide with the state $|\varphi\rangle$ of the system. Still, under the learning rule, the agent will develop a policy where measurement angles α that are closer to φ will be chosen more frequently. From the Born rule, which we used in Eq. (5.26), we can compute the probability that the agent, when measuring O_α , will find the eigenvalue

⁴⁸ Note that the state $|\varphi\rangle$ in Eq. (5.31)—and thus the optimal choice of the measurement observable at time Δt that has $|\varphi\rangle$ as its $+1$ eigenstate—does not depend on any actions of the agent before the preparation of the system in state $|0\rangle_x$. The simple quantum system has no memory, thus there is no need for the agent to remember longer action sequences. To model the learning process of the PS agent, one can therefore neglect glow in the learning rule Eq. (5.33).

+1 and thereby obtain a reward $R = 1$. As shown in Eq. (5.26), this probability is given by $\Pr(\uparrow, \vec{n}_\alpha; \Delta t) \equiv \Pr(+1 | \varphi, \alpha) = |\langle \alpha | \varphi \rangle|^2 = (1 + \cos(\varphi - \alpha)) / 2$.⁴⁹ Therefore, under the policy learnt by the PS agent through its experimenting, it will be rewarded with the probability

$$\Pr_{\text{reward}}^{(t)} = \Pr^{(t)}(+1 | \varphi) = \sum_{\alpha} \Pr(+1 | \varphi, \alpha) p^{(t)}(\alpha) = \sum_{\alpha} \Pr(+1 | \varphi, \alpha) \frac{h_{\alpha}^{(t)}}{\sum_{\alpha'} h_{\alpha'}^{(t)}}. \quad (5.35)$$

This could be interpreted as the agent's individual and embodied probability of correctly guessing the system's state. Similarly, the transition probabilities $p(\alpha)$ can be seen as the agent's embodied discretised probability distribution for estimating the target angle φ , and thus, indirectly, for estimating the magnetic field B .⁵⁰ This distribution develops as the agent interacts with the system and learns how to obtain rewards more often. The agent's behaviour thereby implicitly encodes information about the magnetic field B . The embodied learning of a discrete probability distribution $p(\alpha)$ by the PS agent can be compared with other schemes, such as Bayesian learning, that directly generate a probability distribution $\Pr(\varphi)$ for the angle φ . In that case, one would start with an initially flat distribution $\Pr(\varphi)d\varphi = d\varphi/(2\pi)$ over the range $[0, 2\pi)$. After the first measurement, along one of the directions $\alpha_1 \in \{0, \pi/2, \pi, 3\pi/2\}$ with result $r_1 \in \{+1, -1\}$, the distribution is updated according to Bayes's rule

$$\Pr(\varphi | r_1) = \frac{\Pr(r_1 | \varphi) \Pr(\varphi)}{\Pr(r_1)}. \quad (5.36)$$

Assuming knowledge of the rules of quantum mechanics, i.e., Born's rule $\Pr(r_1 | \varphi) = (1 + r_1 \cos(\varphi - \alpha_1)) / 2$, the terms can easily be calculated. After a series of M measurements, with successive Bayesian updates, one obtains the resulting posterior distribution

$$\Pr(\varphi | r_1, r_2, \dots, r_M) = \frac{1}{\mathcal{N}} \sum_{m=1}^M (1 + r_m \cos(\varphi - \alpha_m)), \quad (5.37)$$

where $\mathcal{N}$ is a corresponding normalisation factor.

Results obtained by Tiersch et al. (2015) indicate that the estimate of the angle φ from the weights h_{α} in the ECM is comparable with the Bayesian method, at least in cases when φ is close to one of the possible choices of α . More generally, the estimates tend to be biased towards the value of α that is closest to φ . It

⁴⁹ Similarly, the probability of obtaining no reward is given by $\Pr(-1 | \varphi, \alpha) = (1 - \cos(\varphi - \alpha)) / 2$.

⁵⁰ This is because $\varphi = \omega \Delta t = |\gamma| B \Delta t$.

should, however, be noted that in contrast to methods such as Bayesian learning or quantum state tomography, there is very little knowledge of quantum physics built into PS. The PS agent does not assume that the rewards originate from measurement probabilities of a quantum state. The PS approach is, in this sense, model free; the PS agent has to build its own implicit model (in the minimal sense of an action-guiding tool as discussed in Sect. 4.4.2), based only on interactions with the environment and the PS learning rule. The goal of the agent is thus not to deliver an explicit estimate of the external magnetic field, but to accumulate positive outcomes of its measurements, by which it implicitly aligns its measurement directions with φ . This alignment dynamics can also deal with time-dependent magnetic fields (as long as these fields change slowly within the time interval Δt), while state estimation methods normally assume stationary environments.

The scenario discussed so far can be generalized in different directions. First, the setting where each measurement round starts with the initial preparation of $|0\rangle_x$ can be straightforwardly generalized to a setting where the agent repeatedly measures the same qubit, at time intervals Δt , without re-initializing it in $|0\rangle_x$. In this setting, every new round starts with the quantum state as updated after the measurement of the preceding round. For example, if, in the first round, the agent chose to measure the observable $O_{\pi/2} = Y$ and obtained the eigenvalue -1 , then the second round starts with $|1\rangle_y$ as the initial state of the qubit. In that setting, the percept that starts a new round is not a simple trigger, but has the general form introduced above, (eigenvalue obtained, measurement direction chosen). This generalised setting is required, for example, in the full scheme of measurement-based quantum computation, where the qubit is part of a larger cluster state and the PS agent has to sequentially adapt different measurement bases. Second, the capacity of the agent to adapt its measurement bases in the presence of the magnetic field is significantly improved with the help of clip composition for the available actions (measurement angles), as this allows the agent to refine its resolution of angles near φ . For more details we refer the reader to the paper by Tiersch et al. (2015).

5.4.4 Design of Photonic Experiments

Most of the applications discussed so far in Sect. 5.4, and also those from Sect. 3.4.4, consider entangled states of qubits. Already in the case of two qubits ($N = 2$), the study of the entanglement properties of the Bell states led to surprising insights, as discussed in the context of Bell's inequalities (Sect. 3.4.3), and to novel applications such as device-independent randomness generation (Sect. 3.5.3), dense coding, and quantum teleportation (Sect. 3.4.4). The generalisation of Bell states to the multi-qubit case ($N > 2$) led to many more examples of entangled states such as GHZ, cluster, and graph states, with further foundational insights and applications such as measurement-based quantum computation. Indeed, in the development of the field of quantum information, the investigation of entanglement has often been driven by a close interplay between foundational questions and practical

applications. Similarly, one can ask which additional insights may be gained when one considers further generalisations towards higher-dimensional systems, that is, entangled states of $N \geq 2$ systems in which each system has a Hilbert space of dimension $d > 2$. From a physics perspective, these considerations are quite natural since many elementary quantum systems have more than two internal states. For example, atoms naturally have a whole series of internal electronic states, trapped chains of ions can be excited into different states of collective oscillatory motion, and molecules exhibit a spectrum of vibrational excitations. The use of such higher-dimensional quantum system, if they can be controlled and manipulated with sufficient precision, provides further possibilities and perspectives for quantum information applications.⁵¹ As a simple example, consider a Bell-type state of two three-dimensional systems (i.e., $N = 2$ and $d = 3$) of the form

$$|\Phi\rangle = \frac{1}{\sqrt{3}} (|00\rangle + |11\rangle + |22\rangle) . \quad (5.38)$$

This state can be used to teleport the state $|\psi\rangle = \sum_{j=0}^2 c_j |j\rangle$ of a three-dimensional quantum system, in a straightforward generalisation of the teleportation protocol of Bennett et al. (1993) presented in Sect. 3.4.4. The generalisation of a graph state ($N \geq 3$) to higher-dimensional systems provides further perspectives and possibilities (see, e.g., Keet et al., 2010). An example for such a state is the generalised GHZ state of N particles in d dimensions. For $N = 3$ and $d = 3$ it has the form

$$|\Phi\rangle = \frac{1}{\sqrt{3}} (|000\rangle + |111\rangle + |222\rangle) . \quad (5.39)$$

While such states can easily be written down formally, the properties and possible applications of multi-partite entangled states in higher dimensions are still little explored and have not been investigated as extensively as for qubits.

A quantum-optical platform for the experimental investigation of such states, and of the entanglement in higher dimensions more generally, are photons in orbital angular momentum (OAM) states (Mair et al., 2001; Malik & Boyd, 2014). These are photons in spatial modes of light that carry orbital angular momentum. Examples are Laguerre-Gaussian modes of a laser beam with a vortex-like behaviour of the phase along the direction of propagation, giving rise to helical wave fronts (Allen et al., 1992; Molina-Terriza et al., 2007). Photons in different OAM modes are then described by states $|m\rangle$ with different orbital angular momenta $\hbar m$ with $m =$

⁵¹ Quantum systems with continuous variables and infinite-dimensional Hilbert spaces add further interesting aspects for the encoding and processing of quantum information (Gottesman et al., 2001). A discussion of these systems is beyond the scope of the current exposition. For a review of discrete and continuous variable systems and their combination in quantum applications, see, e.g., Andersen et al. (2015).

$0, \pm 1, \pm 2, \dots$, where m corresponds to the winding number of the corresponding mode.

The superposition of OAM modes enables one to encode the quantum information of larger quantum systems into the state of a single photon, with interesting perspectives for quantum information processing.⁵² However, the creation of multipartite entangled states of such photons is not straightforward and much less is known than in the two-dimensional case where a qubit is encoded, for example, in a photon's polarisation state. Until recently, little was known on how to entangle three or more OAM photons in a high-dimensional state space and what type of multi-partite entangled states of OAM photons could be created with existing techniques, except in the simplest cases.⁵³ Progress was made by Krenn et al. (2016) who presented a search algorithm (MELVIN) that automatically generated different optical set-ups that were then evaluated by certain objective functions. This led to the finding of set-ups that created novel types of entangled states, and to a survey of different classes of entangled states that could be realised in this way. The specific setting discussed by Krenn et al. (2016) used an optical table with four modes, which were initially excited into a product state of two 3D Bell states, as described in Eq. (5.40), and then transformed via the optical set-up generated by the program. Tripartite entanglement between three of the four photons is generated by putting a detector into one arm of the set-up and post-selecting on the measurement outcome. In other words, a photo-detection event in that selected arm triggers the creation of a three-photon entangled state between the photons in the other three arms.⁵⁴ The question was which type of three-partite entangled states could be created by such a set-up. Between different rounds of iteration of the search algorithm, the results were post-processed in order to simplify the optical set-up and remove redundancies. The mentioned study represented one of the first examples of computer-aided design of novel quantum optical experiments. It also clearly demonstrated the use and power of search algorithms in a situation where the systems under investigation simply become too complex for humans to find a set-up based on previous experience.

The entangled states that can be generated in these experiments are naturally classified by their Schmidt-rank vector (SRV) (r_a, r_b, r_c) . The components of the SRV are natural numbers, denoting the ranks $r_j = \text{rank } \rho_j$ of the reduced quantum states ρ_j (see Eq. (3.105)) of the photons in the corresponding arms $j \in \{a, b, c\}$.⁵⁵

⁵² Examples are discussed, e.g., by Molina-Terriza et al. (2007) and Malik and Boyd (2014). For an early result, see Cerf et al. (2002).

⁵³ The experimental situation is well-described in Malik et al. (2016) and Babazadeh et al. (2017).

⁵⁴ This procedure is necessary since none of the (linear) optical elements in the agent's toolbox can create entanglement in OAM degrees of freedom.

⁵⁵ To connect the Schmidt rank with some examples that we met in Section 3.4.4 consider first the state $|\text{GHZ}\rangle_{123} = \frac{1}{\sqrt{2}}(|000\rangle_{123} + |111\rangle_{123})$. This is the GHZ state of three qubits, which also appears in the Shor code in Eq. (3.166). It corresponds to a (three-qubit) graph state shown in Fig. 3.6c for $N = 3$, for which the reduced quantum state for qubit 1, say, is given by $\rho_1 = \text{tr}_{23}(|\text{GHZ}\rangle_{123}\langle\text{GHZ}|) = \frac{1}{2}(|0\rangle_1\langle 0| + |1\rangle_1\langle 1|)$, which has rank $r_1 = 2$. Similar relations hold for

For example, for the 3D GHZ state of Eq. (5.39), the reduced quantum state of every qubit j is described by a density operator (matrix) $\rho_j = \frac{1}{3}(|0\rangle_j\langle 0| + |1\rangle_j\langle 1| + |2\rangle_j\langle 2|)$, which has rank $r_j = 3$. Therefore, the SRV for that state is (3, 3, 3).

Melnikov et al. (2018b) investigated the application of the PS framework to the same task environment. The goal of this investigation was, similarly, to design quantum experiments that generate different types of high-dimensional entangled OAM states of light. In the PS framework, a basic *action* of the agent consists in placing an optical element, for example a beam splitter or a mirror, on a table, while a *percept* encodes a specific arrangement of optical elements. The optical table is initialized with two entanglement sources⁵⁶ that generate two pairs of entangled photons in the state

$$|\psi_0\rangle_{abcd} = \frac{1}{3} \left(\sum_{m=-1}^1 |m\rangle_a | -m\rangle_b \right) \otimes \left(\sum_{m=-1}^1 |m\rangle_c | -m\rangle_d \right). \quad (5.40)$$

Here the indices a, b, c, d refer to four different arms of the optical arrangement and $|m\rangle_j$ describes the presence of a photon with orbital angular momentum $m\hbar$ ($m = -1, 0, +1$) in arm j . The agent has access to a set of basic optical elements, which includes beam splitters (BS), holograms (Holo), Dove prisms (DP), and mirrors (Refl). Each of these elements realizes a certain unitary transformation on the orbital angular momentum states of one or two photons. For example, a beam splitter BS_{jk} placed at the intersection of two arms, j and k , may transmit a photon from the (incoming) arm j into (outgoing) arm k , or reflect it back into (outgoing) arm j :⁵⁷ $|m\rangle_j \rightarrow (|m\rangle_k + i|-m\rangle_j)/\sqrt{2}$. A Dove prism DP_j placed in arm j multiplies the angular momentum quantum state of the photon by a phase factor depending on m and changes the sign of its angular momentum value: $|m\rangle_j \rightarrow ie^{i\pi m}|-m\rangle_j$, while a mirror Refl_j has a similar effect with a constant overall phase factor of i : $|m\rangle_j \rightarrow i|-m\rangle_j$. Finally, a hologram $\text{Holo}_{j,n}$ in arm j shifts the angular momentum of the photon by the value n : $|m\rangle_j \rightarrow |m+n\rangle_j$. For practical reasons, one may restrict the values of n ; we let $|n| \leq 2$.

The set of basic optical elements, together with their parameters—the choice in which of the arms (or intersections of two arms) the element is placed, and n for the hologram—defines the basic action space of the agent. In our case, the agent is

$j = 2, 3$. This is true for any graph state of N qubits as we demonstrated in Sect. 3.4.4.3 in the derivation leading to Eq. (3.140). The SRV for the GHZ state is thus given by (2, 2, 2). As a second example, consider the state $|\psi\rangle_1|\Phi^+\rangle_{23}$, which appeared as the initial state in the teleportation protocol in Eq. (3.128). It is a tensor product of a pure state $|\psi\rangle_1$ of qubit 1 and the Bell state $|\Phi^+\rangle_{23}$ of qubits 2 and 3. Correspondingly, the reduced states of qubits 2 and 3 are again fully mixed (see the calculation in Eq. (3.107)), while the reduced state of qubit 1 is simply the projector onto the pure state $|\psi\rangle_1$, which has rank $r_1 = 1$. The SRV for that state is therefore (1, 2, 2).

⁵⁶ These sources use lasers and nonlinear crystals in a process called spontaneous parametric down conversion (Malik et al., 2016).

⁵⁷ For the labelling of arms that is used in these experiments, see the example in Fig. 5.11b, where $j = b$ and $k = c$ refer to both the incoming and outgoing arms of the beamsplitter BS_{bc} .

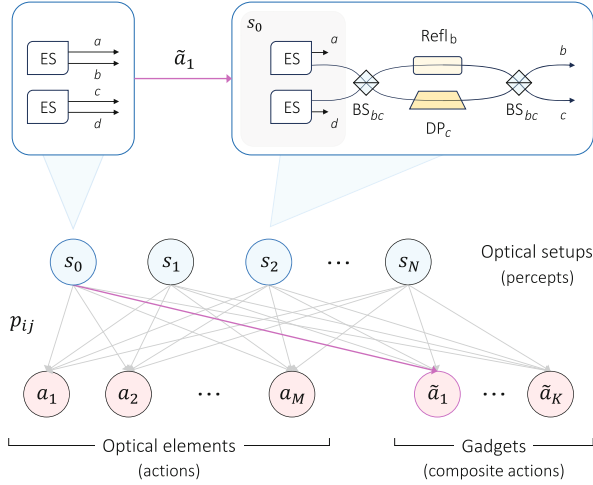


Fig. 5.10 Clip network of a PS agent creating optical experiments. Percept clips (s_i , blue) represent different optical set-ups encountered in the course of creating an experiment. Action clips (a_j , red) represent the placement of an optical element on the table. Every placement of an optical element (a_m) leads to a new set-up, and therefore the percept space grows by adding new elements. Each simulation starts with the percept clip s_0 , corresponding to the empty table, with only the two entanglement sources (ES) mounted, ready to feed two pairs of entangled photons—described by Eq. (5.40)—into the arms a , b , and c , d , respectively. In the lower right part of the figure one can also see gadget clips ($\tilde{a}_k$), corresponding to composite actions that add a whole collection of elements to the table in a single step. The gadget clip $\tilde{a}_1$, for example, represents an interferometer that was composed by the agent in a previous run. The interferometer combines two beam splitters, a mirror, and a Dove prism (see top right inset). It can, for example, be added to the empty table (s_0) in a single step, as indicated by the pink arrow

provided with $|A| = 30$ different basic actions. By placing optical elements on the table, one after the other, the agent successively builds up an experiment.

As already mentioned, a percept represents the state of the optical table, that is, the specific arrangement of the optical elements at a given time. Thus any action of the agent, which consists in adding an optical element to an existing set-up, generates deterministically a new percept that corresponds to the resulting new state of the optical table. The percept space thereby grows dynamically by adding new elements. The corresponding clip network for this task environment is shown in Fig. 5.10. Percept clips are thus representations of previously encountered experimental set-ups, while action clips represent placements of individual optical elements on the table. In this application Melnikov et al. (2018b) also considered a simple form of clip composition, namely action composition, where a composite action consists in placing an entire group of optical elements on the optical table.⁵⁸

⁵⁸ If a sequence of actions results in a rewarded state, then this sequence is added to the agent's repertoire as a single composite action.

Clips that represent composite actions are continuously added to the episodic memory of the PS agent as it finds new, rewarded experiments, that is, experiments that lead to an entangled state of the desired SRV class. Thus both the action and the percept space grow as the agent learns. To prevent overflow of memory, both percept and (composite) action clips can be deleted under rules that depend, roughly speaking, on the reward a resulting set-up has received (for percept clips) and on the strengths of incoming connections to an action clip in the ECM network (for action clips); see Melnikov et al. (2018b). Each time a new element is placed on the (simulated) table, it is evaluated whether entanglement of the desired SRV class has been created.⁵⁹ A reward is issued if the state belongs to a certain entanglement class specified by the SRV (r_a, r_b, r_c) , which depends on the task. This leads to an enhancement of the h -values of the edges shown in Fig. 5.10 proportional to their glow status when the reward is issued.

Melnikov et al. (2018b) consider finite experiments, for which the maximum number of optical elements in one experiment is limited to between six and twelve. This is motivated by both computational and practical restrictions.⁶⁰ The experiment is classified as *interesting* when a reward is obtained, and as *not interesting* when the maximum number of elements is reached without obtaining the reward. The key intuition why a reinforcement learning approach may at all provide an advantage over automated search is that, structurally, the optical table problem (like the error-correcting code-optimisation problem discussed in Sect. 5.4.1) bears some similarity with *grid world*. The main differences are that the world of the agent in the optical-table problem is described by percept and action spaces of variable sizes, and that this world's topology does not correspond to a simple grid, but to a complex network, as illustrated in Fig. 5.11a.

When Melnikov et al. (2018b) applied projective simulation to the optical table problem, the PS agent was given two different types of tasks. The first type of task was to find the simplest set-up that produces some quantum state with a certain Schmidt-rank vector. A specific example was to create an entangled state from the $(3, 3, 2)$ class, which had previously been studied by Malik et al. (2016). After having solved this task, a follow-up task was then to create an entangled state from class $(3, 3, 3)$. The PS agent was not only able to find states from the specified class $(3, 3, 2)$ but it autonomously produced compact set-ups. The built-in optimisation regarding the size of an experiment is enabled by reinforcement learning, as in *grid world*, where the agent is rewarded such as to find the shortest path to a specific site (see Sect. 5.2.4). Remarkably, having found a compact solution to the first task, the

⁵⁹ This includes a post-selection of those cases for which there is exactly one photon in each output mode. In a simulated experiment, assuming that there is no noise or other imperfections, such that one may use pure states, one can compute the exact form of the state and its entanglement at the output, as long as the number of photons and the complexity of the set-up is moderate. In a real experiment, one can certify the creation of a specific state, e.g., by measuring certain entanglement witnesses (Friis et al., 2019).

⁶⁰ For long experiments, one expects a decreasing overall fidelity and loss of entanglement of the resulting state due to an accumulation of imperfections such as slight misalignment.

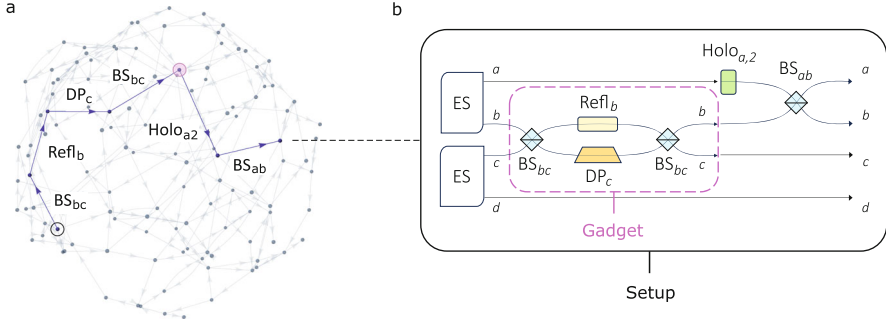


Fig. 5.11 Optical set-ups created by the PS agent. **(a)** Network representing (a small section of) the set of experiments explored by PS. Every vertex represents a specific optical set-up encountered, while every edge represents the placement of an optical element. The creation of an experiment starts with the empty table that contains only the entanglement sources, indicated by a black circle. Highlighted in blue is a sequence of placements that leads from the empty table to the optical set-up shown in panel **(b)**. The filled pink circle on the path indicates a rewarded set-up that contains the gadget highlighted in *pink* in panel **(b)**. The path through the network is generated by corresponding clip-to-clip transitions in the agent’s ECM. If the resulting set-up is rewarded (which depends on the task given to the agent) then a new, composite action clip is created, as detailed in Fig. 5.10

same agent also solved the follow-up task much more efficiently in 50% of all cases, which indicates some form of transfer learning. Overall, the PS agent performed significantly better than automated search and it automatically produced compact experimental set-ups (Melnikov et al., 2018b).

Figure 5.11 shows an example of an optical set-up that was created by the PS agent in the course of the learning process. Different set-ups correspond to vertices in the network shown in Fig. 5.11b, while the edges correspond to specific actions chosen by the agent. The number of different set-ups that can be created grows combinatorially with the number of available actions. If the resulting experiment is rewarded, the corresponding transitions in the ECM that generated that path are enhanced. The path shown in the network of Fig. 5.11a thus maps out the interaction history of the agent, including the elements placed, one after the other, that led to the set-up shown in Fig. 5.11b.⁶¹ The maze shown in Fig. 5.11a suggests an analogy between the task of finding a compact implementation of an interesting experiment and the task of finding the shortest path to a rewarded site in grid world as discussed in Sect. 5.2.4.

The second type of task was to generate as many different experiments as possible that produce high-dimensional multipartite entangled states. Specifically, set-ups were only rewarded if their SRV (r_a, r_b, r_c) satisfied $r_j \geq 2$ ($j = a, b, c$). Furthermore, in order to encourage the exploration of different experiments, the

⁶¹ Note that in that interaction history, the agent did not use the gadget corresponding to the composite action $\tilde{a}_1$.

agent was rewarded only for new findings, i.e., a given set-up was only rewarded once. In that way, the PS agent was used to generate a database of set-ups that create such high-dimensional entangled states, similar as in Krenn et al. (2016). A clear benefit of using PS for this task was that the episodic memory of successful PS agents contained information as to *how* the agent found interesting set-ups, that is, which combinations of actions (sub-set-ups) it often used. This information can be extracted from the structure of the clip network of successful agents. In the PS implementation with the specific form of action composition used by Melnikov et al. (2018b), it is obtained by ranking the memory clips—specifically the clips representing composite actions—according to the sum of the weights of their incident edges.

The top right of Fig. 5.10 displays the same set-up as in Fig. 5.11b, but there it illustrates the role of a *gadget*. A gadget is a combination of optical elements generated by PS that led to a successful experiment (for example, the creation of a state from class (3, 3, 2)). After receiving a reward, which strengthens the connections to the basic action clips (denoted by a_k) that were involved, in addition a new (composite) action clip (here denoted $\tilde{a}_1$) is added to the ECM. Once the gadget is available, the agent can in future rounds create the set-up with a single action. The gadget shown in fact corresponds to an optical interferometer that was previously proposed by Leach et al. (2002) to sort OAM modes with different parities. Thus, the PS agent autonomously (re)discovered a parity sorter in the process of learning how to solve the task of creating high-dimensional entanglement.

It is noteworthy that some gadgets discovered by PS correspond to modern quantum optical devices—designed by humans—for the application in state-of-the-art OAM photonic experiments. Put in slightly more general terms, what has been shown with this analysis is that a PS agent can autonomously create compact set-ups that bring about a certain effect or phenomenon (in this case, high-dimensional entanglement). Gadgets that are generated in that process can be useful beyond the specific given task. Importantly, gadgets are interpretable structures in the PS agent’s episodic and compositional memory.

The study described here constitutes an important step towards autonomous experimenting, but it is still far from genuine scientific explorative research, where we are typically confronted with phenomena that are not yet fully understood. Trying to understand a phenomenon in physics, we usually build a physical model. Such a model should not only explain the observed phenomenon, but also allow us to predict the outcome of yet unperformed experiments. Recent work in this direction is reported by Eva et al. (2023), who investigate the learning of unobserved variables in a three-layered implementation of PS. There it is shown how PS can be used for the development and utilisation of abstract concepts, which could then be employed in the context of scientific experimenting.

5.5 Further Developments of PS and Applications in Other Fields of Research

The applications discussed in Sect. 5.4 focus on quantum information science and technology, targeting technologically fairly advanced task environments. It is expected that in the future more and more tasks will be handled by technological devices that exhibit degrees of artificial agency and learning, including scenarios of fully automated, “self-driving” laboratories (Häse et al., 2019; Rapp et al., 2024). The applications in Sect. 5.4 illustrate how artificial agency modelled by PS can be integrated in quantum laboratories. The main purpose of these studies connects to our book’s subtitle, namely to provide philosophical perspectives on the role of agency in quantum physics. In addition, by modelling scientific agency and by designing artificial scientific agents we may also better understand how we, as human agents, do science, thereby helping us to elucidate our place in the larger course of events in nature. This binds back to the discussion at the very beginning of Chap. 2. The modelling of scientific agency in quantum physics thus also promises to provide more conceptual clarity on how agency integrates with our fundamental theories of nature.

Also in terms of scientific applications, the scope of PS for agency modelling reaches beyond quantum mechanics. In the following we highlight a number of recent developments in PS that address both practical and theoretical issues and which connect to our discussions in Chaps. 4 and 6. This overview is selective; we leave by the wayside a number of further developments of PS, for example, in robotics⁶² and in behavioural biology.⁶³

5.5.1 Large Percept and Action Spaces

When the sizes of the percept and action spaces increase, it takes longer for a PS agent to learn to deal with its environment successfully. Basically, the agent needs to find out the reward for every percept-action pair, which means that, in the worst case, it has to try out all actions. This is a problem in complex environments, which is exacerbated when the agent forgets between different learning rounds, as it is the case for PS ($\gamma > 0$). This problem is also known in standard reinforcement learning, where a table of Q-values is updated according to Eq. (5.17), one value in every round (namely $Q(s^{(t)}, a^{(t)})$ for the specific state-action pair experienced in round t). To deal with the situation of large or even continuous state and action spaces, function approximation methods have been developed. These work with parametrised value-functions that are simultaneously updated over the entire state-action space, including state-action pairs that have not been experienced. Function approximation

⁶² See, e.g., Hangl et al. (2016) and Hangl et al. (2020).

⁶³ See, e.g., Ried et al. (2019), López-Incera et al. (2020, 2021).

based on deep neural networks (in combination with other techniques such as “experience replay”) were presented by Mnih et al. (2015), and they have quickly become standard in value-based reinforcement learning (so-called deep-Q learning). As was shown by Jerbi et al. (2021b), similar function-approximation methods can also be combined with PS, at least for two-layered clip networks, to simultaneously update all transition weights (h -values) of the ECM in every learning round. The specific focus of Jerbi et al. (2021b) was on the treatment of large action spaces and the use of quantum enhancements (see Sect. 5.6), both of which become relevant, e.g., in the context of large-scale quantum experiments. At the same time, this extension of PS adds a further perspective for agency modelling, as it combines transparent with intransparent elements. This becomes relevant for our discussion of explainability and of the specific relation of artificial agency and explainable AI (Sect. 5.7).

5.5.2 *The Use of Multi-layered ECMs*

An important aspect of episodic memory is its dynamic flexibility. As mention in Sect. 5.2, under learning both the weights of the clip network and its topology may change: Clips can be added to and removed from the ECM, and new clips can be composed from existing clips. Examples of such topological changes were given in Sect. 5.4, e.g., in the modelling of photonic experiments (Sect. 5.4.4). There, a new percept clip is added whenever the agent encounters a new set-up, and new action clips are created for gadgets, which are fixed combinations of several optical elements. The corresponding (gadget) clips thus represent entire sequences of actions, each of which may itself already be represented by separate basic action clips. This approach gives the agent much flexibility in representing the environment and its interactions, but the study by Melnikov et al. (2018b) mentioned above is still limited to two-layered ECMs.

Multi-layered and more general clip networks allow one to model more advanced scenarios where the memory develops further structure—based on the agent’s interactions with the environment—that is helpful both for learning and decision finding. Melnikov et al. (2017) discuss an example of a multi-layer clip network where intermediate layers contain clips that represent abstractions from basic percept clips. Such abstractions enable the agent to develop a policy that depends only on selected features of the perceptual input that are relevant for its task environment. In the case of the invasion game that we discussed in Sect. 5.2.4, one can easily consider a more general environment where the perceived signs have more structure. For example, one can think of arrows that not only have different directions, as illustrated in Fig. 5.3, but also different colours, corresponding to different percept categories as introduced at the beginning of Sect. 5.2. In general, the reward may depend on all the structural details, but in many environments only certain aspects of the perceptual input may be relevant—for example, only the shape or only the colour of an object. PS can account for such situations by dynamically generating abstraction clips for which certain features (e.g., the colour or shape)

are masked. As shown by Melnikov et al. (2017), such a mechanism allows the agent to learn even in an environment where no percept appears twice in the agent's interaction history. In the invasion game example this means, for example, that the agent encounters arrows that appear with infinitely many different colours. To deal successfully with such a situation, the agent has to rely on similarities between the different objects rather than repeated occurrences of a given object. The mechanism of clip composition enables the PS agent to solve this task without the need of importing external machinery such as function approximation.

In another study, Eva et al. (2023) propose a three-layered ECM that highlights the model-building role of the ECM in the PS framework. In the setting considered in that work, the percept clips correspond to a set of different objects that the agent can encounter and with which it can conduct a number of different experiments, such as measuring the extension of a spring to which the object is attached in the gravitational field or the motional deflection of the object in a magnetic field. It is assumed that the behaviour of a given object in all the different experiments is completely determined by the values of a set of variables associated with the objects that are unknown to the agent ("hidden variables"), such as mass and net electric charge. An action clip in that study corresponds to an experiment that the agent can perform, together with a prediction of its outcome. An intermediate layer of clips mediates between the layers of percept and action clips, which have no direct connections otherwise. Different from the percept and the action layer, the clips in the intermediate layer do not have a predefined semantics. In a reinforcement learning task, for any encountered object, the agent chooses an action, i.e., an experiment together with a prediction of its outcome. If the experiment is performed and the prediction turns out to be correct, the deliberation path between the percept clip (object encountered) and the action clip (experiment and prediction of outcome), which goes through some clip in the intermediate layer, is enhanced through the standard PS learning rule. As shown by Eva et al. (2023), through the presence of the intermediate layer the agent not only learns to make correct predictions for each object and its associated experiments. At the same time, the pattern of enhanced h -values of transitions in the ECM that develops while the agent learns—both between the percept and the middle layer, and between the middle and the action layer—identifies sets of intermediate clips with different variables. In a sense, these sets then represent embodied abstract concepts associated with the hidden variables. They constitute an action-guiding resource that develops under learning, which helps the agent to understand its environment, as discussed in Sects. 4.5 and 6.6.2.

Multi-layered and general ECMs are also essential for modelling more advanced scenarios of deliberation and decision making, as already discussed in Sect. 5.2.4 (see also Müller & Briegel, 2018). Recent and ongoing work considers various extensions of the PS model that allow one to model deliberations as so-called chains-of-thought. One such extension (LeMaitre et al., 2024) implements deliberation as a random walk of *multiple excitations* in the clip network, which allows for a more flexible way of representing complex thoughts and disentangling them into separate components (concepts) that come together and interact during a deliberation. The

transition probabilities in that approach may depend on the full configuration of excitations, which can be accommodated by using hypergraphs (instead of graphs) for the mathematical description of the ECM. This represents a versatile resource for modelling. Another extension introduces *domain knowledge* that is stored in separate memory registers linked to the ECM and which can be accessed during a deliberation. The integration of domain knowledge provides domain-specific semantics for the deliberations as a random walk and enlarges the interpretability of the PS model (see Sect. 5.7). Multi-layered ECMs have also been used to study the combination of projective simulation with principles from active inference (Linson et al., 2018), as proposed by Pazem et al. (2024).

5.5.3 Quantum Generalisations

A promising feature of the PS model that has been in view from the earliest works on is that the framework offers a natural route for its generalisation into the quantum domain. This is mainly due to the description of deliberations as a random walks, which have known quantum analogues. The question how agency can be integrated with quantum physics is a central topic of this book, as already discussed in Sec. 4.1. We therefore devote a separate section to quantum extensions of the PS model.

5.6 Quantum Extensions of PS

In this section we will discuss quantum extensions of the PS model that employ further quantum resources such as coherence and entanglement—beyond and above the basic agential resource of indeterminism—for specific capacities including decision finding and learning. These extensions are related to and in fact contribute to current efforts for developing quantum-enhanced learning agents.⁶⁴

In Sect. 4.1, we already discussed different approaches to integrating agency with quantum physics. The approach that we adopt in this section belongs to the class (ii) of Sect. 4.1, which considers quantum resources for enhancing certain agential capacities. It does not belong to the class (i) that considers a full quantisation of the agent as a physical system, including its environment with which it interacts. The latter approach can still be meaningful when investigating the overall scope of quantum machine learning (Dunjko et al., 2016) and the different types of quantum speed-ups that can—or cannot—be obtained. This may include the possibility to

⁶⁴ Most of these efforts are discussed under the term *quantum machine learning* (Schuld et al., 2015; Biamonte et al., 2017; Arunachalam & de Wolf, 2017; Dunjko & Briegel, 2018), which encompasses a broad range of research questions with a focus on data-driven approaches, similar as in classical machine learning. Recently, quantum-enhanced methods for reinforcement learning, which come with a basic notion of agency via the reinforcement learning part, have received more attention (Dunjko et al., 2017; Chen et al., 2020; Jerbi et al., 2021a).

access the environment “in superposition”, which requires a quantum mechanical treatment of the environment *and* the agent. For illustration, one can think of a quantum computer interacting with an environment consisting of nodes in a quantum network which, in the extreme case, is described by a fully coherent and unitary quantum dynamics. This may be interesting for new types of quantum information processing, but a (quantum) computer is not an agent and it is not clear how one could reasonably speak of learning in such a scenario.⁶⁵ As argued in Sect. 2.3, without irreversibility, there is no ontological emergence and thus there is no agency. To arrive at a description of the agent proper, one needs to identify an entity that really acts and thereby integrates indeterminism for its capacities as an agent. We touched upon the problems of a fully quantum view of agency already at the beginning of Chap. 4. The discussion of agential possibilities and the role of indeterminism as a key feature of agency will be taken up again in Sects. 6.3 and 6.5.

5.6.1 Quantum Resources for Deliberation and Learning

As introduced in Sect. 5.2, the PS model is based on stochastic processes, in particular the random walk that underlies the deliberation dynamics. In so far as the clip-to-clip transitions in the ECM are assumed to be truly random, quantum mechanics may already be needed as a resource for indeterminism. One could thus argue that PS, as a physical model, already assumes quantum mechanics at the basic level. Indeterministic quantum processes could be realized, e.g., in optical implementations by placing photodetectors behind (tunable) beam splitters and sampling from the measurement outcomes.

The idea of quantum resources for a PS agent, however, goes further than the assumption of quantum indeterminism as a source for true randomness. As we have seen in Sect. 3.4.4, quantum enhancements based on quantum coherence and entanglement are exploited in various applications of quantum information and computation.⁶⁶ For a PS agent, such quantum enhancements can arise when it is endowed with a quantum memory. This means that different clips can be excited in a quantum superposition and the deliberation dynamics is realised by

⁶⁵ In the classical domain, notions of learning are usually based on the interaction (that is, perception) *history* of the agent with the environment: “I saw this, then I did that, but that was bad, so I won’t do it again . . .”. To learn from past experiences, one needs to know *what happened*, for which there need to be real events in the first place. This suggests that for learning, and thus for agency, irreversibility is essential. If nothing really happens, nothing can be learnt. These considerations connect to our discussion of indeterminism and the ontological emergence of agency in Sect. 2.3.

⁶⁶ To mention just one proposal as an example: The wide range of applications that are currently being investigated in the framework of the European Quantum Flagship initiative includes quantum-enhanced sensing and metrology, communication, simulation, computation, and basic science (European Union, 2024).

coherent quantum walks. The agent then has the capacity, in a sense, to recall past experiences in superposition. The basic intuition here derives from an analogy of the clip network with an interferometer. In an optical interferometer, as discussed in Sect. 3.3.3.5, light can travel along different paths through an arrangement of beamsplitters and mirrors, which gives rise to the phenomenon of interference. The intensity of light at a given output port of the interferometer depends on all the paths light can take to arrive at this point. Light waves travelling along different paths can interfere constructively or destructively, that is, their amplitudes add up to enhance or diminish the total signal, respectively. If the light beam is replaced by a single photon travelling through the interferometer, the resulting phenomenon is called quantum interference. Instead of the *intensity* of the light beam at the output, one then considers the *probability* that a photon is detected at a certain output port. The photon, as it were, travels along different paths of the interferometer in superposition, and therefore a change anywhere in the interferometer (e.g., putting some optical medium or phase shifter in one of the arms) will typically change the detection probabilities at all output ports. The optical interferometer is analogous to a quantum-enhanced clip network: Different modes of the interferometer correspond to different clips in the ECM network.

In the following we will briefly discuss three different physical realisations of quantum enhanced PS, each of them following a different approach and focusing on a different aspect of enhancement.

5.6.2 *Speed-up of Deliberation via Quantum Walks*

Since in PS the deliberation process is driven by a random walk, it is natural to ask how one would describe deliberation in a quantum PS agent. A natural generalisation of classical random walks are so-called *quantum walks*. Indeed, the existence of speed-ups of quantum walks over classical random walks on certain types of graphs (Kempe, 2003) provided one early motivation for a quantisation of the PS model. However, these speed-ups were established only for specific, highly symmetric graphs, such as high-dimensional hypercubes. The graph describing a clip network in the PS model is, in contrast, rather generic; it depends on the specific learning history of an individual PS agent and thus changes over time. To exploit the advantages of quantum walks in the context of PS, one needs to connect the PS framework to more generic results on quantum speedups for searching on graphs and to Markov chain-based algorithms that were established by Szegedy (2004) and other groups (for example, Magniez et al., 2011). To make use of these results, the PS framework can be modified in a way that is close in spirit to the idea of reflection in the deliberation process, as introduced in Briegel and De las Cuevas (2012) and as discussed in Sect. 5.2.4.⁶⁷ Deliberation, envisioned as a random walk through the

⁶⁷ This is described and motivated in Paparo et al. (2014).

network of clips, is thereby described via a Markov chain. In the language of Markov chains, the process of deliberation then corresponds to mixing, and the process of action selection corresponds to sampling from the equilibrium distribution. This translation of the standard PS framework to a “reflecting PS” model enabled Paparo et al. (2014) to establish a quadratic speed-up in the decision finding (that is, deliberation and action selection) for a class of ECMs and Markov chains.⁶⁸ This result showed that one can make sense, in a rigorous way, of the intuitive idea of recalling (that is, re-exciting) episodic memory in superposition and exploiting quantum interference for faster decision finding. The mentioned result is based on sophisticated quantum algorithmic techniques such as quantum phase estimation and amplitude amplification (Magniez et al., 2011).

5.6.3 Quantum Many-body Approach to Deliberation

A more physics-motivated approach to quantizing memory and the deliberation process in PS was outlined by Briegel and De las Cuevas (2012), where the episodic memory is described by a quantum many-body Hamiltonian⁶⁹ and deliberation corresponds to the propagation of certain excitations through the memory. Given the classical description of episodic memory as a graph $G = (V, E)$ as in Sect. 5.2.1, each vertex $j \in V$ is associated with a pair of quantum mechanical operators $\hat{c}_j^\dagger$ and $\hat{c}_j$ that describe the creation or annihilation of an excitation of the corresponding clip, respectively. The quantum mechanical state describing the excited (activated) clip j is thus given by $\hat{c}_j^\dagger|\text{vac}\rangle$, where $|\text{vac}\rangle$ denotes the ground state. The quantum memory is then described by a Hamiltonian of the form

$$H = \sum_{(j,k) \in E} (\lambda_{jk} \hat{c}_k^\dagger \hat{c}_j + \lambda_{jk}^* \hat{c}_k \hat{c}_j^\dagger) + \sum_{j \in V} \epsilon_j \hat{c}_j^\dagger \hat{c}_j, \quad (5.41)$$

with couplings $\lambda_{jk} = \lambda_{kj}^*$ and real-valued energies ϵ_j . The term $\lambda_{jk} \hat{c}_k^\dagger \hat{c}_j$ in the first sum describes the propagation (hopping) of an excitation from clip j to clip k , while the term $\epsilon_j \hat{c}_j^\dagger \hat{c}_j$ in the second sum accounts for the energy ϵ_j associated with the excitation of clip j .⁷⁰ The propagation of excitations in the quantum memory is then described by the dynamical (von Neumann) equation, which we already met in Sect. 3.2.3 (Eq. (3.35)):

$$\frac{\partial}{\partial t} \rho = \frac{1}{i\hbar} [H, \rho], \quad (5.42)$$

⁶⁸ See also Dunjko and Briegel (2015) and Orsucci et al. (2018).

⁶⁹ The classical analogue would be a (generally non-linear) system of coupled oscillators, where clips correspond to the excitations of the fundamental oscillatory modes.

⁷⁰ The coupling strengths λ_{jk} can be complex, but it is required that $\lambda_{kj} = \lambda_{jk}^*$ because H must be a Hermitean operator.

where $\rho = \rho(t)$ is the quantum state (density operator) of the memory at time t and, to recall, the commutator $[H, \rho] \equiv H\rho - \rho H$.

The dynamics generated by Eq. (5.42) is reversible, that is, the directedness of the propagation of an excitation is only due to the initial conditions. To be able to retrieve the classical dynamics of deliberation from the quantum model in some limiting sense, one needs a process that generates transitions only in certain directions, for example, from clip j to clip k but not *vice versa*. These directed processes can be modelled by a Lindblad-type generator (super-operator) L of the form

$$L\rho = \sum_{(j,k) \in E} \kappa_{jk} \left(\hat{c}_k \hat{c}_j^\dagger \rho \hat{c}_k^\dagger \hat{c}_j - \frac{1}{2} (\hat{c}_k^\dagger \hat{c}_j \hat{c}_k \hat{c}_j^\dagger \rho + \rho \hat{c}_k^\dagger \hat{c}_j \hat{c}_k \hat{c}_j^\dagger) \right) \quad (5.43)$$

with $\kappa_{jk} \geq 0$, which is an effective description of irreversible processes due to interactions of the memory with a bath of uncontrolled degrees of freedom. When combined with the coherent interactions between the clips, described by the Hamiltonian from Eq. (5.41), one obtains the quantum master equation

$$\frac{\partial}{\partial t} \rho = \frac{1}{i\hbar} [H, \rho] + L\rho, \quad (5.44)$$

which generalizes Eq. (5.42). Equation (5.44) represents a quantum generalization of the master equation that describes the classical random walk, which is recovered in the limit $H = 0$, i.e., when $\lambda_{jk} = 0$ for all $(j, k) \in E$. The coherent transitions generated by the Hamiltonian term, i.e., by $[H, \cdot]$ in Eq. (5.44), can give rise to quantum interference of different paths in the clip network. The transitions generated by the Lindblad term L , in contrast, are incoherent and can be interpreted as the spontaneous quantum jumps between different clip states.

An appealing feature of this Hamiltonian model is that it highlights an important, embodied, feature of episodic memory. Viewed properly, episodic memory is not a passive device for the storage of data, like the random access memory in a computer, but rather incorporates the capacity of re-invoking past experiences. On this view, re-invoking past experiences corresponds to a simple form of remembering, which is an active, dynamic process.

Past experiences of the agent shape the ECM, that is, the topology and the weights of the clip network described by the graph $G = (V, E)$, which in the quantum model contains the couplings of H and L . One could say that, in order to recall past experiences based on a given percept, they always have to be dynamically re-created (run through again, i.e., simulated).

Percepts and actions in this model are described classically. Upon encountering a percept $s \in S$, a corresponding clip $c = \mu(s)$ is activated that is described by the quantum state $\hat{c}^\dagger |\text{vac}\rangle = \hat{\mu}^\dagger(s) |\text{vac}\rangle$. The subsequent propagation of the excitation through the ECM as described by Eq. (5.44) then drives the deliberation, which now can be regarded as a quantum simulation.

It is interesting to consider the situation when the percept and action spaces, S and A , have further structure, as described at the beginning of Sect. 5.2. In the quantum model considered here, this structure corresponds to a tensor-product in the corresponding Hilbert space describing the ECM. In case of a composite percept $s = (s_1, \dots, s_N)$, the corresponding clip operator has the form

$$\begin{aligned}\hat{c}^\dagger &= \hat{\mu}^\dagger(s) \\ &= \hat{\mu}_1^\dagger(s_1) \otimes \hat{\mu}_2^\dagger(s_2) \otimes \dots \otimes \hat{\mu}_N^\dagger(s_N) \\ &= \hat{c}_1^\dagger \otimes \hat{c}_2^\dagger \otimes \dots \otimes \hat{c}_N^\dagger,\end{aligned}\tag{5.45}$$

where $\hat{c}_j^\dagger = \hat{\mu}_j^\dagger$ is a clip operator that excites percept category j , representing, for example, a certain colour or shape.⁷¹ With composite clip operators $\hat{c}^\dagger = \hat{c}_1^\dagger \otimes \hat{c}_2^\dagger \otimes \dots \otimes \hat{c}_N^\dagger$, H in Eq. (5.41) features highly non-trivial many-body interactions which are difficult to realise experimentally in the laboratory. However, quantum devices with engineered quantum many-body interactions have been studied in the context of dissipation-driven quantum simulation, e.g., with trapped ions (Barreiro et al., 2011) or neutral atoms in optical lattices (Weimer et al., 2010). For a recent review of methods and applications of engineered quantum dissipative processes see Harrington et al. (2022).

5.6.4 Quantising PS on a Photonic Chip

In the remainder of this section, we focus on a proposal for a quantum extension of PS that is in several aspects close to the Hamiltonian picture, but offers a concrete physical implementation. In this proposal, recently introduced by Flamini et al. (2024) and subject of ongoing work both theoretical and experimental (Franceschetto & Ricou, 2024; Gaspar et al., 2024), deliberation is also modelled as the quantum walk of physical excitations. These are here realised as photons in an optical interferometer. Different from the Hamiltonian-Lindbladian approach discussed above, however, the directionality of the transitions need not be enforced by dissipation, but can be naturally realised by the geometric layout of the interferometer and the choice of input ports that encode the percept clips. The proposal discusses quantum extensions of the learning dynamics of PS that could be naturally implemented on a photonic chip and that are in part inspired by recent advances in the field of quantum machine learning (Cerezo et al., 2021; Jerbi et al., 2021a).

⁷¹ Note that $\hat{\mu}_j^\dagger(s_j)$ can also be set equal to the identity, depending on the value of s_j . For example, we can set $\hat{\mu}_j^\dagger(\star) = I$ if we choose $\star$ to describe the absence of that feature in the total percept.

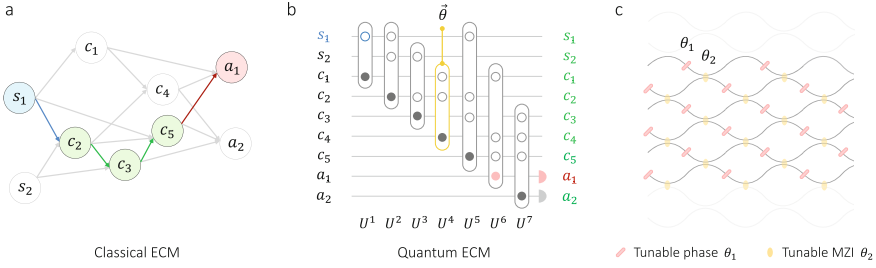


Fig. 5.12 Photonic realisation of an episodic memory. (a) A classical random walk in the ECM; (b) the corresponding quantum walk displayed as a sequence of unitary transformations U^l , each of which acts on a subset of modes corresponding to different clips. (c) An array of Mach-Zehnder interferometers with tunable phase shifters. Such arrays are used to realise the unitary transformations U^l that compose the scattering matrix U^{ECM} of Eq. (5.47). The set of interferometric phases $\vec{\theta}$ is updated under feedback from the environment, which allows the agent to learn. Adapted from Flamini et al. (2024)

Specifically, we consider the realization of the ECM in an array of tunable Mach-Zehnder interferometers as shown in Fig. 5.12c.⁷² A discussion of the basic Mach-Zehnder interferometer was given in Sect. 3.3.3.5. As it allows the coherent manipulation of photonic states, it is a central and versatile tool for optical implementations of quantum information protocols. As shown by Flamini et al. (2024), for any ECM described by a directed acyclic graph (DAG), a quantum version can be obtained which corresponds to a sequence of unitary transformations that act on different subsets of clips at a time. A classical random walk in the ECM thereby becomes a quantum walk of a single excitation through different layers of an interferometric array, each of which implements a unitary transformation U^l on a subset of modes corresponding to different clips (see Fig. 5.12). The classical transition probabilities, described by the set of h -values, are thereby replaced by quantum transition amplitudes, described by the matrix elements of U^l .

In this optical implementation, a clip $c_j \in C$ corresponds to a mode in the interferometric array (Flamini et al., 2020). In the quantum case, we associate with each optical mode a pair of operators $\hat{c}_j^\dagger, \hat{c}_j$, similar as in the Hamiltonian model of Eq. (5.41). For the following, we use the Heisenberg picture, such that the time dependence lies in the mode operators, i.e., $\hat{c}_j^\dagger = \hat{c}_j^\dagger(\tau)$.⁷³ A deliberation process then starts with the excitation of some clip c_j at time τ_0 , which corresponds, in the simplest case, to one of the input modes of the interferometric network in Fig. 5.12c.

⁷² The general idea of this platform was introduced by Reck et al. (1994); it is here used in the form presented by Clements et al. (2016).

⁷³ We use the variable τ for the description of deliberation processes internal to the agent. This variable is different from the external, discrete time variable t , which counts the rounds of agent-environment interactions that drive the agent's learning. In standard reinforcement-learning applications, τ is typically not resolved and considered negligible compared to the time scale of learning. In PS, however, τ is resolved and refers to the time-scale of decision finding.

The subsequent propagation of the excitation is described by a unitary operator of the form

$$\hat{c}_k^\dagger(\tau_n) = U^\dagger(\tau_n, \tau_0) \hat{c}_k^\dagger(\tau_0) U(\tau_n, \tau_0) = \sum_{j \in V} U_{kj}^{\text{ECM}} \hat{c}_j^\dagger(\tau_0), \quad (5.46)$$

which relates the excitation of the modes (clip operators) of the network at different times. The scattering matrix U^{ECM} that appears in Eq. (5.46) is, in turn, composed of a sequence of n intermediate transformations U^l ,

$$U^{\text{ECM}}(\vec{\theta}) = \prod_{l=1}^n U^l(\vec{\theta}), \quad (5.47)$$

corresponding to the different layers l in Fig. 5.12b.

The U^l , and thus the entire scattering operator U^{ECM} in Eq. (5.47) depend on a set of parameters $\vec{\theta}$ that can be controlled by tunable phase shifters (Reck et al., 1994) as shown in Fig. 5.12c. The parameters $\vec{\theta}$ encode the state of the ECM at a given time and thus play a role similar to the h -values of the standard PS model. These parameters are updated under feedback from the environment, i.e., when the agent learns, as described below.

The specification of a complete deliberation process is as follows. Encountering percept $s \in S$, deliberation starts with the excitation of a percept clip ($\hat{c}_s^\dagger$) at time τ_0 , corresponding to the input mode s , and ends with the detection (i.e., de-excitation) of some action clip ($\hat{c}_a$) at time τ_n , corresponding to the output mode a , which, in turn, triggers a corresponding action on the environment. The quantum mechanical transition amplitude for that process is given by

$$\langle \text{vac} | \hat{c}_a(\tau_n) \hat{c}_s^\dagger(\tau_0) | \text{vac} \rangle = \langle a; \tau_n | s; \tau_0 \rangle, \quad (5.48)$$

where we introduce the notation $|c_k; \tau_l\rangle \equiv \hat{c}_k^\dagger(\tau_l) | \text{vac} \rangle$ for the time-dependent eigenstates of the number operators $\hat{c}_k^\dagger(\tau_l) \hat{c}_k(\tau_l)$ corresponding to an excitation being present in clip c_k at time τ_l . The probability $p(a|s)$ to perform action a as a result of the deliberation, after having encountered percept s , is thus given by the modulus square of this expression, that is,

$$p(a|s) = |\langle a|s \rangle|^2 \equiv |\langle a; \tau_n | s; \tau_0 \rangle|^2 = \left| U_{as}^{\text{ECM}}(\vec{\theta}) \right|^2. \quad (5.49)$$

Note that in the short-hand notation used in the first part of Eq. (5.49), the eigenstates $|a\rangle$ and $|s\rangle$ implicitly refer to clip operators at different times. This usage mirrors the notation for the agent's policy, the conditional probability $p(a|s)$, whose arguments also refer to different times τ_0 and τ_n , namely to the beginning and the end of the

deliberation, respectively.⁷⁴ Equation (5.49) is a quantum version of Eq. (5.2), and it entails the quantum interference of all possible paths an excitation can take through the clip network, which can be exploited for decision finding.

Note that the input and output coupling functions $I(c|s)$ and $O(a|c)$ appearing in Eq. (5.2) could easily be accommodated in the quantum case as well. In going from Eq. (5.2) to (5.49) we have considered directed acyclic graphs for $G = (V, E)$ describing episodic memory, which allow one to enforce a definite duration of the deliberation process, corresponding to the n unitaries U^l in the decomposition of the scattering matrix of Eq. (5.47). However, directed graphs including loops can always be represented by sets of directed acyclic graphs. More advanced optical implementations may also include the possibility of routing photons back, which would be another way of realising more general ECMs and to model processes such as *reflection* discussed in Sect. 5.2.

As was shown by Flamini et al. (2024), the proposed optical platform is highly versatile and allows various extensions relevant for PS. For example, one can effectively remove or add new optical modes without the need to modify the hardware, simply by adjusting phases in the mesh of Mach-Zehnder interferometers (see Fig. 5.12c). This is particularly relevant for the implementation of topology-changing features of PS, such as the creation of vertices and edges of the graph $G = (V, E)$ representing the ECM, and for clip composition.

Thus far we have discussed the quantum optical implementation of the episodic memory as a physical structure for deliberation. Another important feature of PS is the realisation of learning by way of enhancing transitions in the clip network that led to rewarded actions. In the classical version of PS, the deliberation process that connects perceptual input with action output maps out a single specific path through the clip network. In the quantum version, this is no longer the case: Deliberation is here driven by a quantum coherent process, where different clips are excited in parallel, i.e., in a quantum superposition. There are various possibilities for enabling learning in such a fundamentally different situation. One approach is inspired by quantum machine learning and the use of parametrised quantum circuits for reinforcement learning (Jerbi et al., 2021a); a different approach is based on a direct update of the scattering operator U^{ECM} in Eq. (5.47) describing the ECM. In both cases, the set of parameters $\vec{\theta}$ that specify U^{ECM} , implemented by a specific setting of the phase shifters in Fig. 5.12c, are updated in a way that is based on the reward.

While there is no unique way to generalise Eq. (5.5) to the quantum domain, Flamini et al. (2024) motivate the following learning rule:

$$p_{sa}^{(t+1)} = \frac{1}{|A|} + (1 - \gamma) \left(p_{sa}^{(t)} - \frac{1}{|A|} \right) + r^{(t)} g_{sa}^{(t)}, \quad (5.50)$$

⁷⁴ This is fully consistent with the Heisenberg picture that we adopt here, whereas in the Schrödinger picture this may be confusing. Indeed, if $|a\rangle$ and $|s\rangle$ as used in Eq. (5.49) would refer to the same time, they would be orthogonal, and then the expression $\langle a|s\rangle$ in Eq. (5.49) would make no sense.

where $p_{sa}^{(t+1)} \equiv p_{sa}(\vec{\theta}^{(t+1)})$ and $p_{sa}^{(t)} \equiv p_{sa}(\vec{\theta}^{(t)})$, and $r^{(t)}$ is a (rescaled) reward for the quantum PS agent. Note that Eq. (5.50) directly involves the transition probabilities between percept and action clips—essentially, the policy. Operationally, this means that under the reward $r^{(t)}$, the parameters $\vec{\theta}$ of the ECM need to be changed in such a way that the new policy $p_{sa}^{(t+1)}$ relates to the old policy $p_{sa}^{(t)}$ as specified in Eq. (5.50). Flamini et al. (2024) consider different methods for updating the policy; these either work directly on the parameters $\vec{\theta}$ (phases shifter settings) by a variational procedure based, e.g., on gradient descent to minimize a loss function corresponding to Eq. (5.50); or they work on the matrix elements of the scattering operator U^{ECM} which determines the policy via Eq. (5.49). For a detailed exposition of these methods including some fine print, their learning performance, and how they can be optimized under experimental constraints, we refer the reader to Flamini et al. (2024).

The optical implementation of quantum PS can be developed in various directions, including the possibility to change the topology of the clip network under learning by the deletion and creation of clips and their connections, as already mentioned. It also provides a straightforward way of implementing multi-excitation walks, where the number of clips that are excited at any given time τ_l can be larger than one. The theory of multi-excitation PS is still under development even in the classical case, but it is clear, for example, that this extension will enable conditional random walks that can be used for more complex forms of deliberation and reasoning.⁷⁵ Multi-excitation PS also provides a natural framework for the representation of composite clips, which, in the quantum case, involves tensor products of excitation operators similar as in Eq. (5.45) in the Hamiltonian model. Thus the optical implementation proposed by Flamini et al. (2024) and experimentally demonstrated by Franceschetto and Ricou (2024) represents a versatile platform for a realization of quantum projective simulation (“Q-PS”).

For our philosophical discussion, we should stress that this proposal maps out, in concrete detail, a physical model of an artificial learning agent whose central agential capacities—deliberation and learning—are described by quantum mechanics. As pointed out at the beginning of this section, this is still work in progress. It is an open question how precisely to understand the model-building capacity of an agent, which we discussed in Chap. 2 and specifically in Sect. 4.5, in quantum PS. First steps connecting the quantum platform to the idea of concept formation in multi-layered classical PS (see Eva et al. (2023) and Sects. 5.5 and 6.6.2) are outlined in Flamini et al. (2024). At any rate, these investigations can be seen as a proof of principle that an *implicit model* arising under learning—in the sense of an action-guiding tool as discussed in Sect. 4.4.2—may well be quantum. In Sect. 5.7, we will use this implementation of Q-PS in a discussion of the transparency of PS and artificial agency more generally, and its relation to the topics of interpretability and explainability of AI.

⁷⁵ A recent development in this area is reported by LeMaitre et al. (2024).

5.7 Interpretability of PS

In current efforts to develop trustworthy AI systems, notions of transparency, interpretability and explainability are being discussed extensively in the field of *explainable artificial intelligence* (XAI). As we have shown in Sects. 5.5 and 5.4.4 (see also Sect. 2.7.1), the structures and physical processes described by the ECM play a central role for the interpretation of PS as an agency model. When we discuss AI in terms of artificial agency, the PS model will allow us to discuss explainability and relate it to *traceability* as a fundamental physical property.⁷⁶

To illustrate what is meant by explainability, it is perhaps best to start with a negative example. Current AI systems that are based on deep neural networks are often *black box* systems. In the context of classification, one may ask *why* a given input led to a specific output: For example, what was it about the input that led to its misclassification? A black box will not give a satisfactory answer.

An example for the input could be the perception of a certain traffic situation, where the misclassification could result in an inappropriate or even fatal behaviour of a self-driving car. To get an idea what we expect as an explanation, a possible strategy is to imagine that, in the same situation, the misclassification happened to a human driver. What would we accept as an explanation from the driver? The situation is perhaps clearer in a context where the AI is not used for classification, but involves some kind of reasoning. In Sect. 5.2.4, we showed how PS can be used as a basic model for reasoning. The random walk in ECM is then understood as a progression of considerations, which is triggered by a given perceptual situation and ends with some action output.

5.7.1 Explanations in Classical PS

The pertinent features of classical PS that are relevant for the issue of explainability are all connected to its episodic memory (ECM) and the mechanisms of learning and deliberation. These features are, specifically:⁷⁷

1. the semantics of clips that derives from, and builds upon, the basic meaning of percepts and actions;
2. the process of deliberation as a random walk that maps out a specific trajectory (deliberation path) in the clip network; and
3. the process of learning through rewards, which leads to the growth of the *h*-values of specific transitions in the clip network.

⁷⁶ In Chap. 6 we will adopt the position that a deeper understanding of the scope of artificial intelligence requires us to understand AI as artificial agency (Dreyfus, 1992).

⁷⁷ As we will discuss in what follows, features (2) and (3) are indeed specific for classical, non-quantum PS, and (3) can fail when classical PS is enhanced by neural networks.

With these features, PS provides an instance of a transparent and explainable AI system (Gunning et al., 2019, 2021): (i) The deliberation path explains how the agent, encountering a percept, came to a decision or action; (ii) the deliberation process can be traced with the help of a simple record, i.e., a log file; (iii) the explanation can be easily communicated by the machine by providing the log file, possibly together with the description of the clip network.

Point (iii) depends on the kind of questions we ask and highlights the issue of what could be called a long term *vs.* short term explanation. The actual deliberation path provides an explanation in the sense of a sequence of considerations (or chain of thoughts) that connects the percept encountered with the action taken. The structure of the clip network provides an explanation in the sense of which considerations *could* have come to the agent's mind, i.e., it maps out the space of available considerations, together with the weights that correspond, one could argue, to the agent's dispositions for reasoning in specific ways. The actual learning history—which is not explicitly stored in the ECM, but which shapes its topology and the structure of its weights—also explains the dispositions.

In PS, explaining is thus grounded in retracing a physical process, which is enabled by the structure of the ECM as a network of clips.⁷⁸ This notion of explanation makes sense in a wide range of contexts. When we interact with a PS agent, it can communicate its deliberation path as its explanation for how it arrived at a specific decision in a given situation. Similarly, the manufacturer of an artificial PS agent can give an account of the behaviour of its product to a customer. In case of malfunction or unacceptable behaviour, the manufacturer can do something about it. They can, for example, re-educate the agent by exposing it to certain environments that facilitate the reconfiguration of its ECM; alternatively, they could directly operate on the ECM, for example by deleting or adding clips and edges (Flamini et al., 2024). The PS agent, viewed as an AI system, thus exhibits at least some of the main elements of a trustworthy AI.⁷⁹

Explanation takes on a slightly different meaning when a PS agent is viewed as an artefact *vs.* as an agent. On the first view, the manufacturer explains the behaviour of its product. On the second view, however, the agent can explain itself, which seems to entail a certain degree of transparency and communication. These two ways of thinking about explanation naturally come together when we talk about artificial agents.

It is important to realise that the possibility of *tracing* the physical processes that underlie the dynamics of an AI system does not necessarily enable an explanation of its decisions. A counterexample is provided by deep neural networks, as mentioned at the beginning of the section. When looking for an explanation for a given instance

⁷⁸ This is true as long as the clips retain a simple semantics. We will shortly describe cases in which this property may be lost.

⁷⁹ This list includes the points 4. *Transparency. Including traceability, explainability and communication.* and 7. *Accountability. Including auditability, etc.* from the Ethics Guidelines for Trustworthy AI issued by the European Commission (High-Level Expert Group on Artificial Intelligence, 2019).

of misclassification, it won't help to trace the micro-state of the neural network, that is, the set of its weights and how they evolved during training, even though this could in principle be done. On the other hand, while tracing of the physical micro-state of the AI system may not be sufficient for an explanation, there are also examples where it is not necessary, as will be illustrated next.

5.7.2 *Human Agents Giving Explanations*

For illustration, consider the example of a defendant, that is, a *human agent* who has to testify in court. If asked by the judge to explain why he acted in a specific way, the defendant is supposed to give reasons for his actions in the specific situation. These reasons may be quite informal, i.e., the defendant may simply tell his story, how he experienced the situation, including a description of his emotional state. Such a statement of experience could be taken as the defendant's complete explanation of his behaviour. At any rate, that is the explanation the judge and the jury will initially be dealing with, considering whether it is authentic, credible, and so on.⁸⁰ To qualify as an explanation in that sense, it is not essential that the statements given by the defendant be correct—the statements of experience made by the defendant may not correspond to what he really experienced at the time (after all, he could remember incorrectly), and his report may not correspond to what actually happened (he may have perceived the situation in a way that does not match with the reported experience of other witnesses).

Importantly and relatedly, when asking the defendant for an explanation of his way of acting, the judge will not ask for a physical record of the firings of all neurons in the defendant's brain while he was acting in the specific situation. This information, even if it were available, would normally be useless for the jury. First, they would typically not know how to interpret it, and second, it would not count as an explanation given by the agent in his capacity as an agent. However, information of that type could still become relevant when it is noticed that the defendant consistently argues in some strange way, hallucinates, or when there are reasons to believe that he may suffer from a medical condition that affects his memory or mental experience in general. In such cases, it would be justified, or even called for, to take the agent's verbal explanations with care. If some organic problem is diagnosed, that might limit which parts of the interrogation the judge can use in the trial.

Interrogation is based on communication with the defendant, which addresses him *as a human agent* and typically attributes his actions in the human-specific,

⁸⁰ Of course, based on this initial report, the defendant in a trial may need to answer further questions of the judge, who might be challenging him to reflect on and evaluate his behaviour, which typically involves moral considerations. This part will not be applicable if we consider lower level, non-human and therefore non-moral agents.

moral way. As discussed in Sect. 2.3.2, besides human-specific moral and general agential attribution, there is also mere causal attribution, which can consider the agent from a medical or biological point of view. If merely causal attribution becomes pertinent, focus shifts and the agent (and often specifically his brain) will be viewed *as a physical and biological system*. This shift exemplifies Strawson's so-called *objective attitude* mentioned in Sect. 2.3.2 in the context of rung 3. An explanation of the agent's behaviour on that view can be wholly dissociated from what the agent may have said.

All the different types of attribution can be used to ground explanations, but these operate at different levels. As human agents are also physical systems, they can be the target of different types of attribution—the whole ladder of attribution described in Sect. 2.3.2 is potentially available. Normal human explanations for their behaviour, prompted by a question like “Why did you do that?”, pertain to the highest, human-specific rung of the attribution ladder and thus can connect with reflective considerations of morality. Explanation can, however, also address other forms of attribution, down to a description of brain states. In our view, it is highly important in the debate on explainable AI to be aware of these different levels of explanation and not to mix them up.

5.7.3 Transparency and Intransparency in Classical PS

The causal level of attribution, which in the human case corresponds to the description of bodily or brain states at a given time, in the PS model corresponds to the description of the record of deliberation for a given round. This record is provided by the trace of the random walk that drives the deliberation and is recorded in the glow matrix. If that random walk has a direct physical implementation, the explanation for an agent's behaviour at that lowest level is directly provided by the record of the physical process that drives the deliberation.

However, this simple connection may get lost if the ECM and its physical realisation are more complex. As an example, consider a situation with very large percept and action spaces where learning in the standard PS model becomes too slow. As described in Sect. 5.5, one method to accelerate the learning process is to integrate PS with a deep neural network. So-called deep PS (Jerbi et al., 2021b) is an example of a model that has both an intransparent and a transparent component. The intransparent part drives the process of learning, where the updates of the ECM's weights are realised by a deep neural network. The transparent part describes the process of deliberation, which is realised by the random walk through the ECM. Deep PS is thus an example where the explainability of *learning* may get lost, despite the in-principle traceability of the physical processes in the deep neural network. At the same time, the explainability of the agent's *decision finding* is retained, due to the traceability of the random walk in the semantically conspicuous ECM.

A second example in which explainability of PS comes under pressure is when the structural evolution of the ECM erodes the semantics of the clip network. This could occur when many new clips with no predefined semantics are created and added to the network. If the resulting structure becomes too complex and does not satisfy further constraints such as modularity, even the explainability of deliberation, that is, the short-time aspects can be affected. Without a semantics for the ECM, tracing the deliberation path of the random walk through the clip network will no longer be sufficient for an explanation.

5.7.4 Explainability of Quantum PS

Quantum versions of PS come with further challenges regarding explainability. This is relevant both for theoretical considerations regarding the physical foundations and underpinnings of the model, and for practical considerations regarding technological realisations and quantum enhancements of learning agents.

The main issue with all quantum versions of PS is that, due to the fundamental wave-particle duality, one cannot retain full information about the deliberation process in the ECM: Traceable deliberation paths as in classical PS are impossible. The ramifications of this for the explainability of a PS agent's decisions need to be investigated in detail. For concreteness, we focus on the photonic realisation of PS discussed in Sect. 5.6.4, and we discuss the short-term and the long-term dynamics separately.

First, regarding the decision-making (i.e., short-time) dynamics: In photonic Q-PS, the classical random walk is replaced by a quantum walk, which corresponds to the propagation of a photon through the interferometer representing the clip network. On its way through the interferometer the photon leaves no trace. After an action was performed by the agent, one therefore cannot answer the question which deliberation path was taken, i.e., how exactly the perceptual input was coupled with the action output. This ability would violate the quantum mechanical principle of wave-particle duality, which we discussed in Sect. 3.3.3.5 in the context of the Mach-Zehnder interferometer. In the present context, Englert's duality relation (Eq. (3.89)) expresses a trade-off between distinguishability, which is the basis of traceability, and the capacity of quantum interference, which is the basis of any quantum speed-up. Generalisations of the mentioned duality relation exist (Dürr, 2001; Bagan et al., 2018), but so far there are no results for quantum walks on interferometer networks with arbitrary topology (e.g., including loops). If such further generalisations could be established, one might be able to derive fundamental, quantitative relations between the achievable efficiency (quantum speed-up) and the explainability (in the sense of traceability) of the quantum agent's decision-making process.

Second, regarding the learning (i.e., long-term) dynamics: As discussed in Sect. 5.6, since no definite path can be credited for a rewarded decision-making process, the reward cannot be assigned exactly as in the classical model. For this

reason, Flamini et al. (2024) introduced a learning rule based on a variational procedure that affects the weights of multiple paths simultaneously. In the classical model, the changes of the weights in the clip network can be traced over long time scales, which can explain how the dispositions of the agent are shaped through its learning history. As mentioned earlier, such an explanation could also be obtained from a (hypothetical) record of all transitions that have been taken in the ECM during the agent's history: Classically, the weights (h -values) in the ECM encode the relative frequency of such transitions. Quantum mechanically, due to the aforementioned wave-particle duality, such a record cannot exist. However, one can still obtain some information on which transitions in the interferometric array of Fig. 5.12c realizing the ECM are likely to be visited during a deliberation: one can in principle compute the probabilities of detecting a photon in the different modes of the interferometer along the deliberation process, i.e., at different layers. In a hypothetical situation where the agent is queried and interrupted during its deliberation, these would be the probabilities with which the agent would report, as it were, its current consideration (i.e., which clip in its ECM is currently excited). It is not yet clear how helpful this kind of information would be—in particular for explanations regarding the development of the agent's dispositions—and how it could be leveraged to enable straightforward explainability as in the classical, plain PS model. But it is certainly worthwhile to explore traceability along these lines.

Third and finally, we note that explainability in Q-PS may be achieved without the full traceability that is offered by the classical PS model. Classically, the random walk maps out a specific sequence of considerations that constitutes the process of deliberation and which afterwards explains the action. Quantum mechanically, one could argue that the agent weighs all considerations that are present in its memory and chooses an action as a result of the entire weighing process. It is not clear whether that second view must be less convincing, or explain less. On the negative side, one can argue that an explanation must be a report of precisely which considerations came to the agent's mind at any given time or stage of the deliberation process; it must be a report of the exact sequence of thoughts that occurred to the agent. But positively, one can argue that an explanation (in a weaker but still useful sense) is already constituted by an account of all considerations that are somehow present and which played a role in the decision finding process. The direct classical correspondence of this second view would be to identify the explanation with the entire stochastic process. Note that this does not amount to replacing the entire decision-finding process with a stochastic mapping of inputs (percepts) to outputs (actions). Instead, the explanation would lie in the unravelling of the stochastic map into a specific random walk (in the classical case), or in the unravelling of the unitary map into a specific quantum walk (in the quantum case).

Many of these considerations and ideas, especially in the quantum domain, but also in the classical domain where PS is combined with other methodologies such as neural networks, are still part of ongoing research. In any case, these considerations illustrate the intriguing interplay between the transparency and traceability of

physical processes constituting the agent's memory and the explainability of the agent's actions.

5.8 Conclusion on PS

In the final Section of this Chapter, we take a higher level perspective and consider PS with a view to its capacity for model-building, as discussed in Chap. 4, and as an agency model, in relation to our discussion from Chap. 2. This discussion also provides additional motivation for our consideration of Phenomenology in Chap. 6.

5.8.1 *Modelling Model-building Agents: The Perspective of PS*

The issues of transparency and explainability are closely connected to the problem of agency modelling, and specifically to the modelling of scientific agency.⁸¹ As argued in Sect. 4.5, modelling scientific agency amounts to modelling the agent's capacity for modelling. The meaning of modelling is itself a topic in the philosophy of science (Frigg & Hartmann, 2025), as discussed in Sect. 4.4. Our use of this term here aligns with the idea of models as tools, as argued in Sect. 4.4.2. Models are theoretical tools insofar as they summarise and distil the experience of an agent in its dealing with its environment and, at the same time, help it to use past experiences in new situations.

The basic features of a PS agent—projection, simulation, and learning—are all grounded in its memory structure, i.e., in its ECM: Past experiences of the agent are re-invoked in a random walk through the ECM to prepare an action, and the ECM itself changes under new experiences. Similarly, models in PS must be understood as specific structures in the ECM that have two particular defining features: they develop under learning, and they enhance the agent's capacity of projection.

In the study mentioned in Sect. 5.5.2, Eva et al. (2023) considered a minimal learning agent with a three-layered ECM that learns how to make correct predictions while experimenting with different objects. While the agent learns, clips in an intermediate layer of its ECM become connected to the other layers in such a way that they acquire the role of abstract concepts, or variables. In this study, those variables correspond to mass, size and charge of the given objects. The intermediate layer of clips can then be used as an “interface through which PS agents

⁸¹ When modelling scientific agency, we imagine an agent that performs experiments on some system of interest, which is part of its environment. The motivation (or goal) of the agent is to *understand* the behaviour of the system. In scientific practice, understanding is developed by building a good model of the system that provides intelligible explanations (De Regt, 2017). A good model should also enable the agent to correctly predict the behaviour of the system in novel experiments that have not been performed yet.

can naturally articulate and communicate explanations of their reasoning processes and decisions” (Eva et al., 2023, 215). At the same time, this layer can be used successfully by the agent to generalize and predict the outcome of yet unperformed experiments. The resulting structures exhibit at least some of the main features of physical models.

The closeness to physical models distinguishes a PS agent’s modelling from the modelling considered in so-called model-based reinforcement learning. As discussed in Sect. 5.3, the environment of a reinforcement learning agent is often described as a Markov Decision Process, and the transition function $T(s, a, s')$ of that process (Eq. (5.13)) is referred to as the *model of the environment*. Model-based reinforcement learning agents estimate that function, for example as a look-up table, in order to solve the given task environment. By way of contrast, the PS agent’s models, as structures in the ECM, are much closer to models used in physics. Physical models should provide a compact description of the behaviour of the system (mathematical or otherwise), not just a mere look-up table in which function values $T(s, a, s')$ for all input triples are listed.⁸²

The PS approach to modelling scientific agents as modellers sketched here is still more of a research program. An important next step would be to study the PS agent’s modelling for at least the first two of the three levels of difficulty of the general modelling loop discussed in Sect. 4.4. A major further challenge would then be to tackle the third level of difficulty, i.e., to get PS to exhibit real scientific creativity. The concrete task could be, for example, to find a model for the Zeeman effect and then also for the anomalous Zeeman effect (see Sect. 3.1.2). In order to enable the PS agent to solve this task, it could first be given learning environments in which it can familiarise itself with the effect of magnetic fields on moving charges, with the effect of currents on magnetic dipoles, and with emission spectra of various chemical elements. In a next step, these environments could be combined. Once the agent has learnt to predict the normal Zeeman effect, it could then be confronted with the surprising phenomenon of an even splitting of spectral lines that constitutes the anomalous Zeeman effect.

While that has to be left to future work, the general understanding of models as structures in the ECM already provides some interesting perspectives for QBism that relate to our discussion in Sect. 4.5. Suppose, first, that we use a PS agent *as a device* that is trained in a set of experiments in order to develop a model of the system under study and eventually to output predictions (in the form of a probability distribution) for the outcome of yet unperformed experiments. In the classical case, such experiments could, for example, involve billiard balls whose trajectory the agent should predict, while in the quantum case they could involve the prediction of an electron’s position in the ground state of some potential. In the classical case, it seems clear that the probability distribution learnt by the device

⁸² Similarly, we would not consider a deep neural network trained as a function approximator, or some other black-box program that merely outputs $T(s, a, s')$ for every input triple (s, a, s') , as a *physical* model of the system.

will have some physical correspondence (or embodiment) in its ECM. For example, the deliberation dynamics can be imagined as a simple mechanical scattering process that generates the probability distribution $p(a, \tau|s, \tau_0)$.⁸³ Inspecting the PS device's memory would then allow one to identify the mechanism that, when run, implements the policy: the memory would constitute an embodied version of the device's policy. In the quantum case, the corresponding structure could be the Hamiltonian of Eq. (5.41), or the mesh of Mach-Zehnder interferometers in Fig. 5.12. The corresponding quantum scattering processes generate the probability distribution $p(a, \tau|s, \tau_0) = |\langle a, \tau|s, \tau_0 \rangle|^2$, with the transition amplitude $\langle a, \tau|s, \tau_0 \rangle$ given by Eq. (5.49).

Suppose now that we view the PS agent not as a device for us, but *as an agent*. In QBism an agent is an entity that has its own state of belief about the environment, i.e., about the system it is experimenting with. The PS agent's state of belief, and thus the probabilistic predictions it makes, will naturally depend on its individual learning history. These predictions can therefore be different for individual agents with different learning histories. From this perspective it is obvious that the PS agent's probability distribution will not be a description of the experimental system as such, but of the agent's accumulated experience in dealing with it. In that respect, this view complies with one of the main claims of QBism, namely that the wave function and thus the probability distribution that an agent assigns to a system is personal to the agent, something like its personal state of belief about the system. But in contrast with QBism, in which the agent is not further resolved and is just a peg to hang that belief state on, the PS approach shows that the agent's state of belief is embodied in the state of its ECM. PS agents build models that are themselves physical structures in their memories. This in turn supports our general realism about agency. And it also offers a useful perspective on the intersubjective nature of science that QBism does not resolve either. Different PS agents will build different models even when interacting with the same system. For a deeper understanding of the system and for more accurate predictions of the system's behaviour it will in general be useful to compare and combine these different models. This corresponds well with our general view of science as a collaborative social enterprise that we mentioned in Sect. 2.6.3.

5.8.2 The Status of PS as an Agency Model

In this book we have argued that agency can be characterised by its capacity for projection and simulation, and in this chapter, we have described the agency model of Projective Simulation together with some of its applications. Taking a step back

⁸³ One could think of a scattering process implemented via a system of pipelines with different cross sections that terminate at the action-clips, similar to a Galton board. On that view, one can argue that Galton's board embodies the binomial distribution.

from the details of our discussion, we can now characterise the status of PS as an agency model by referring back to the four criteria (i)–(iv) for agency that we developed in Sect. 2.6.1.

- i. *Embodiment.* PS as defined mathematically in Sect. 5.2 is a formal model, but the percept and action spaces, S and A , already point toward embodied realisations. Any such realisation of PS will have to use real, material sensors and actuators, and it will have to rely on some concrete physical hardware that implements the ECM. Embodied realisations of PS agents are subject of ongoing research. The quantum-optical demonstration discussed in Sect. 5.6.4, however, already serves as a proof-of-principle.
- ii. *Settling, indeterminism, and agency-incompatibilism.* PS is indeterministic at root, and the decision finding process of a PS agent amounts to settling the status of its actuators, which is not a deterministic function of the sensory input. In an embodied PS agent, these actuators can plausibly be regarded as parts of its body. In this sense, a PS agent really settles matters regarding its body, and this supports the thesis of agency-incompatibilism. With respect to the next criterion, note that an entity that has no open possibilities that it can settle ceases to be an agent.
- iii. *Wanting.* In Chap. 2 we identified *wanting* as a complex condition on agency whose fulfilment is not easy to check. When an agent acts such as to collect rewards, its behaviour will typically be identified as goal-directed. (Recall Anscombe's (1957, §36) remark that "[t]he primitive sign of wanting is *trying to get*".) The deeper question is whether the agent *truly wants* the rewards or is just programmed to collect them, which would be a defeater (see Sect. 2.4). As we will underline via our phenomenological analysis of agency in Chap. 6, however, there is a deeper form of wanting that is conceptually tied to the very notion of agency: An agent wants open possibilities that it can meaningfully settle, because at root, being an agent *is* having open possibilities for the future. PS agents align with this criterion because they are always projecting into the open future, and they maintain open possibilities through adaptation.
- iv. *Historicity, learning, and individuality as diachronic aspects of agency.* The central structure of a PS agent is its ECM, its Episodic and Compositional Memory, which grounds a notion of historicity. As an agent's decision finding is fundamentally indeterministic, the concrete structure of an agent's ECM develops individually: even two PS agents interacting with the same environment will typically end up with different ECMs. The diachronic development of an agent is fundamentally tied to its individual learning history, and this leads to individual behavioural dispositions.

In summary, PS is a rich model for agency that shows specific promise as a model for scientific agency. It is also well aligned with the criteria for agency that we established via our explication in Sect. 2.6.1. As we will see in the next chapter, the PS model also affords a phenomenological interpretation, which further strengthens its systematic appeal.

References

- Allen, L., Beijersbergen, M. W., Spreeuw, R. J. C., & Woerdman, J. P. (1992). Orbital angular momentum of light and the transformation of Laguerre-Gaussian laser modes. *Physical Review A*, 45, 8185–8189.
- Andersen, U. L., Neergaard-Nielsen, J. S., van Loock, P., & Furusawa, A. (2015). Hybrid discrete- and continuous-variable quantum information. *Nature Physics*, 11(9), 713–719.
- Anscombe, G. E. M. (1957). *Intention*. Blackwell, Oxford. 2nd edition 1963, reprint Harvard University Press, Cambridge, MA, 2000.
- Arunachalam, S., & de Wolf, R. (2017). Guest column: A survey of quantum learning theory. *SIGACT News*, 48(2), 41–67.
- Aslam, N., Zhou, H., Urbach, E. K., Turner, M. J., Walsworth, R. L., Lukin, M. D., & Park, H. (2023). Quantum sensors for biomedical applications. *Nature Reviews Physics*, 5(3), 157–169.
- Babazadeh, A., Erhard, M., Wang, F., Malik, M., Nouroozi, R., Krenn, M., & Zeilinger, A. (2017). High-dimensional single-photon quantum gates: Concepts and experiments. *Physical Review Letters*, 119, 180510.
- Bagan, E., Calsamiglia, J., Bergou, J. A., & Hillery, M. (2018). A generalized wave-particle duality relation for finite groups. *Journal of Physics A: Mathematical and Theoretical*, 51(41), 414015.
- Barreiro, J. T., Müller, M., Schindler, P., Nigg, D., Monz, T., Chwalla, M., Hennrich, M., Roos, C. F., Zoller, P., & Blatt, R. (2011). An open-system quantum simulator with trapped ions. *Nature*, 470(7335), 486–491.
- Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press.
- Bennett, C. H., Brassard, G., Crépeau, C., Jozsa, R., Peres, A., & Wootters, W. K. (1993). Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters*, 70, 1895–1899.
- Bennett, C. H., Brassard, G., Popescu, S., Schumacher, B., Smolin, J. A., & Wootters, W. K. (1996). Purification of noisy entanglement and faithful teleportation via noisy channels. *Physical Review Letters*, 76, 722–725.
- Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., & Lloyd, S. (2017). Quantum machine learning. *Nature*, 549, 195–202.
- Bjerland, Ø. F. (2015). Projective simulation compared to reinforcement learning. Master's thesis, Department of Informatics, University of Bergen. <https://hdl.handle.net/1956/10391>.
- Bohg, J., Hausman, K., Sankaran, B., Brock, O., Kragic, D., Schaal, S., & Sukhatme, G. S. (2017). Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33, 1273–1291.
- Boyajian, W. L., Clausen, J., Trenkwalder, L. M., Dunjko, V., & Briegel, H. J. (2020). On the convergence of projective-simulation-based reinforcement learning in Markov decision processes. *Quantum Machine Intelligence*, 2(2), 13.
- Bravyi, S., Cross, A. W., Gambetta, J. M., Maslov, D., Rall, P., & Yoder, T. J. (2024). High-threshold and low-overhead fault-tolerant quantum memory. *Nature*, 627, 778–782.
- Briegel, H. J. (2001). Principles of entanglement purification. In D. Bouwmeester, A. Ekert, A. & Zeilinger (Eds.), *The physics of quantum information: Quantum cryptography, quantum teleportation, quantum computation* (pp. 261–268). Springer.
- Briegel, H. J. (2012). On creative machines and the physical origins of freedom. *Scientific Reports*, 2, 522.
- Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, 2, 400.
- Briegel, H.-J., Dür, W., Cirac, J. I., & Zoller, P. (1998). Quantum repeaters: The role of imperfect local operations in quantum communication. *Physical Review Letters*, 81, 5932–5935.
- Briegel, K. D., von Grafenstein, N. R., Draeger, J. C., Blümmler, P., Allert, R. D., & Bucher, D. B. (2025). Optical widefield nuclear magnetic resonance microscopy. *Nature Communications*, 16(1), 1281.

- Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., Fujii, K., McClean, J. R., Mitarai, K., Yuan, X., Cincio, L., & Coles, P. J. (2021). Variational quantum algorithms. *Nature Reviews Physics*, 3(9), 625–644.
- Cerf, N. J., Bourennane, M., Karlsson, A., & Gisin, N. (2002). Security of quantum key distribution using d -level systems. *Physical Review Letters*, 88, 127902.
- Chen, S. Y.-C., Yang, C.-H. H., Qi, J., Chen, P.-Y., Ma, X., & Goan, H.-S. (2020). Variational quantum circuits for deep reinforcement learning. *IEEE Access*, 8, 141007–141024.
- Clements, W. R., Humphreys, P. C., Metcalf, B. J., Kolthammer, W. S., & Walmsley, I. A. (2016). Optimal design for universal multiport interferometers. *Optica*, 3(12), 1460–1465.
- De Regt, H. W. (2017). *Understanding scientific understanding*. Oxford Studies in Philosophy of Science. Oxford University Press.
- Delfosse, N., & Nickerson, N. H. (2021). Almost-linear time decoding algorithm for topological codes. *Quantum*, 5, 595.
- Dennis, E., Kitaev, A., Landahl, A., & Preskill, J. (2002). Topological quantum memory. *Journal of Mathematical Physics*, 43(9), 4452–4505.
- Deutsch, D., Ekert, A., Jozsa, R., Macchiavello, C., Popescu, S., & Sanpera, A. (1996). Quantum privacy amplification and the security of quantum cryptography over noisy channels. *Physical Review Letters*, 77, 2818–2821.
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. The MIT Press.
- Dunjko, V., & Briegel, H. J. (2015). Quantum mixing of Markov chains for special distributions. *New Journal of Physics*, 17(7), 073004.
- Dunjko, V., & Briegel, H. J. (2018). Machine learning & artificial intelligence in the quantum domain: A review of recent progress. *Reports on Progress in Physics*, 81, 074001.
- Dunjko, V., Taylor, J. M., & Briegel, H. J. (2016). Quantum-enhanced machine learning. *Physical Review Letters*, 117, 130501.
- Dunjko, V., Taylor, J. M., & Briegel, H. J. (2017). Advances in quantum reinforcement learning. In *2017 IEEE international conference on systems, man, and cybernetics (SMC)* (pp. 282–287).
- Dürr, S. (2001). Quantitative wave-particle duality in multibeam interferometers. *Physical Review A*, 64, 042113.
- Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17, 449–467.
- Erhard, A., Poulsen Nautrup, H., Meth, M., Postler, L., Stricker, R., Stadler, M., Negnevitsky, V., Ringbauer, M., Schindler, P., Briegel, H. J., Blatt, R., Friis, N., & Monz, T. (2021). Entangling logical qubits with lattice surgery. *Nature*, 589(7841), 220–224.
- European Union (2024). Quantum flagship projects. <https://qt.eu/projects/>.
- Eva, B., Ried, K., Müller, T., & Briegel, H. J. (2023). How a minimal learning agent can infer the existence of unobserved variables in a complex environment. *Minds and Machines*, 33(1), 185–219.
- Flamini, F., Hamann, A., Jerbi, S., Trenkwalder, L. M., Poulsen Nautrup, H., & Briegel, H. J. (2020). Photonic architecture for reinforcement learning. *New Journal of Physics*, 22(4), 045002.
- Flamini, F., Krumm, M., Fiderer, L. J., Müller, T., & Briegel, H. J. (2024). Towards interpretable quantum machine learning via single-photon quantum walks. *Quantum Science and Technology*, 9, 045011.
- Fowler, A. G., Mariantoni, M., Martinis, J. M., & Cleland, A. N. (2012a). Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86, 032324.
- Fowler, A. G., Whiteside, A. C., & Hollenberg, L. C. L. (2012b). Towards practical classical processing for the surface code. *Physical Review Letters*, 108, 180501.
- Franceschetto, G., & Ricou, A. (2024). Demonstration of quantum projective simulation on a single-photon-based quantum computer. *Physical Review A*, 110, 062613.
- Frigg, R., & Hartmann, S. (2025). Models in science. In E. N. Zalta (ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, summer 2025 edition. <https://plato.stanford.edu/archives/sum2025/entries/models-science/>.

- Friis, N., Vitagliano, G., Malik, M., & Huber, M. (2019). Entanglement certification from theory to experiment. *Nature Reviews Physics*, 1(1), 72–87.
- Gaspar, J. M., Bergerault, A., Apostolou, V., & Ricou, A. (2024). Entanglement-enhanced quantum reinforcement learning: an application using single-photons. In *2024 IEEE international conference on quantum computing and engineering (QCE)* (Vo. 02, pp. 329–334).
- Glesener, G. (2020). Adaptive measurement-based quantum computation using projective simulation in the presence of graph state symmetries. Master's thesis, University of Innsbruck.
- Gottesman, D., Kitaev, A., & Preskill, J. (2001). Encoding a qubit in an oscillator. *Physical Review A*, 64, 012310.
- Grassl, M. (2023). New quantum codes from CSS codes. *Quantum Information Processing*, 22(1), 86.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120.
- Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). Darpa's explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4), e61.
- Hangl, S., Ugur, E., Szedmak, S., & Piater, J. (2016). Robotic playing for hierarchical complex skill learning. In *Intelligent robots and systems (IROS), 2016 IEEE/RSJ international conference* (pp. 2799–2804).
- Hangl, S., Dunjko, V., Briegel, H. J., & Piater, J. (2020). Skill learning by autonomous robotic playing using active learning and exploratory behavior composition. *Frontiers in Robotics and AI*, 7, 42.
- Harrington, P. M., Mueller, E. J., & Murch, K. W. (2022). Engineered dissipation for quantum information science. *Nature Reviews Physics*, 4(10), 660–671.
- Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1, 282–291.
- High-Level Expert Group on Artificial Intelligence (2019). Ethics Guidelines For Trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>.
- Jerbi, S., Gyurik, C., Marshall, S., Briegel, H., & Dunjko, V. (2021a). Parametrized quantum policies for reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. Wortman Vaughan (Eds.), *Advances in neural information processing systems* (Vol. 34, pp. 28362–28375). Neural Information Processing Systems Foundation.
- Jerbi, S., Trenkwalder, L. M., Poulsen Nautrup, H., Briegel, H. J., & Dunjko, V. (2021b). Quantum enhancements for deep reinforcement learning in large spaces. *Physical Review X Quantum*, 2, 010328.
- Joo, J., Alba, E., García-Ripoll, J. J., & Spiller, T. P. (2013). Generating and verifying graph states for fault-tolerant topological measurement-based quantum computing in two-dimensional optical lattices. *Physical Review A*, 88, 012328.
- Kaelbling, L. P. (1994). Associative reinforcement learning: Functions ink-DNF. *Machine Learning*, 15(3), 279–298.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4(1), 237–285.
- Keet, A., Fortescue, B., Markham, D., & Sanders, B. C. (2010). Quantum secret sharing with qudit graph states. *Physical Review A*, 82, 062315.
- Kempe, J. (2003). Quantum random walks: An introductory overview. *Contemporary Physics*, 44(4), 307–327.
- Krenn, M., Malik, M., Fickler, R., Lapkiewicz, R., & Zeilinger, A. (2016). Automated search for new quantum experiments. *Physical Review Letters*, 116, 090405.
- Krinner, S., Lacroix, N., Remm, A., Di Paolo, A., Genois, E., Leroux, C., Hellings, C., Lazar, S., Swiadek, F., Herrmann, J., Norris, G. J., Andersen, C. K., Müller, M., Blais, A., Eichler, C., & Wallraff, A. (2022). Realizing repeated quantum error correction in a distance-three surface code. *Nature*, 605(7911), 669–674.
- Langford, J., & Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20). Neural Information Processing Systems Foundation.

- Leach, J., Padgett, M. J., Barnett, S. M., Franke-Arnold, S., & Courtial, J. (2002). Measuring the orbital angular momentum of a single photon. *Physical Review Letters*, 88, 257901.
- LeMaitre, P. A., Krumm, M., & Briegel, H. J. (2024). Multi-excitation projective simulation with a many-body physics inspired inductive bias. arXiv 2402.10192.
- Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2), 279–311.
- Linson, A., Clark, A., Ramamoorthy, S., & Friston, K. (2018). The active inference approach to ecological perception: General information dynamics for natural and artificial embodied cognition. *Frontiers in Robotics and AI*, 5, 21.
- López-Incera, A., Ried, K., Müller, T., & Briegel, H. J. (2020). Development of swarm behavior in artificial learning agents that adapt to different foraging environments. *PLoS One*, 15(12), e0243628.
- López-Incera, A., Nouvian, M., Ried, K., Müller, T., & Briegel, H. J. (2021). Honeybee communication during collective defence is shaped by predation. *BMC Biology*, 19(1), 106.
- Magniez, F., Nayak, A., Roland, J., & Santha, M. (2011). Search via quantum walk. *SIAM Journal on Computing*, 40(1), 142–164.
- Mair, A., Vaziri, A., Weihs, G., & Zeilinger, A. (2001). Entanglement of the orbital angular momentum states of photons. *Nature*, 412(6844), 313–316.
- Malik, M., & Boyd, R. W. (2014). Quantum imaging technologies. *La Rivista del Nuovo Cimento*, 37(5), 273–332.
- Malik, M., Erhard, M., Huber, M., Krenn, M., Fickler, R., & Zeilinger, A. (2016). Multi-photon entanglement in high dimensions. *Nature Photonics*, 10(4), 248–252.
- Mautner, J., Makmal, A., Manzano, D., Tiersch, M., & Briegel, H. J. (2015). Projective simulation for classical learning agents: A comprehensive investigation. *New Generation Computing*, 33, 69–114.
- Melnikov, A. A., Makmal, A., & Briegel, H. J. (2018a). Benchmarking projective simulation in navigation problems. *IEEE Access*, 6, 64639–64648.
- Melnikov, A. A., Poulsen Nautrup, H., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., & Briegel, H. J. (2018b). Active learning machine learns to create new quantum experiments. *Proceedings of the National Academy of Sciences*, 115, 1221–1226.
- Melnikov, A. A., Makmal, A., Dunjko, V., & Briegel, H. J. (2017). Projective simulation with generalization. *Scientific Reports*, 7(14430).
- Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1), 8–30.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.
- Molina-Terriza, G., Torres, J. P., & Torner, L. (2007). Twisted photons. *Nature Physics*, 3(5), 305–310.
- Müller, T., & Briegel, H. J. (2018). A stochastic process model for free agency under indeterminism. *Dialectica*, 72(2), 219–252.
- Nigg, D., Müller, M., Martinez, E. A., Schindler, P., Hennrich, M., Monz, T., Martin-Delgado, M. A., & Blatt, R. (2014). Quantum computations on a topologically encoded qubit. *Science*, 345(6194), 302–305.
- Orsucci, D., Briegel, H. J., & Dunjko, V. (2018). Faster quantum mixing for slowly evolving sequences of Markov chains. *Quantum*, 2, 105.
- Paparo, G. D., Dunjko, V., Makmal, A., Martin-Delgado, M. A., & Briegel, H. J. (2014). Quantum speedup for active learning agents. *Physical Review X*, 4(3), 031002.
- Pazem, J., Krumm, M., Vining, A. Q., Fiderer, L. J., & Briegel, H. J. (2024). Free energy projective simulation (feps), Active inference with interpretability. arXiv 2411.14991.
- Popescu, S. (1994). Bell's inequalities versus teleportation: What is nonlocality? *Physical Review Letters*, 72, 797–799.
- Poulsen Nautrup, H., Delfosse, N., Dunjko, V., Briegel, H. J., & Friis, N. (2019). Optimizing quantum error correction codes with reinforcement learning. *Quantum*, 3, 215.

- Preskill, J. (2018). Quantum computing in the NISQ era and beyond. *Quantum*, 2, 79.
- Rapp, J. T., Bremer, B. J., & Romero, P. A. (2024). Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1), 97–107.
- Raussendorf, R., & Briegel, H. J. (2001). A one-way quantum computer. *Physical Review Letters*, 86, 5188–5191.
- Reck, M., Zeilinger, A., Bernstein, H. J., & Bertani, P. (1994). Experimental realization of any discrete unitary operator. *Physical Review Letters*, 73, 58–61.
- Ried, K., Müller, T., & Briegel, H. J. (2019). Modelling collective motion based on the principle of agency: General framework and the case of marching locusts. *PLoS ONE*, 14(2), 1–21.
- Russell, S., & Norvig, P. (2022). *Artificial Intelligence: A Modern Approach* (4th ed.). Pearson Education.
- Schuld, M., Sinayskiy, I., & Petruccione, F. (2015). An introduction to quantum machine learning. *Contemporary Physics*, 56(2), 172–185.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
- Szegedy, M. (2004). Quantum speed-up of Markov chain based algorithms. In *45th annual IEEE symposium on foundations of computer science* (pp. 32–41).
- Tiersch, M., Ganahl, E., & Briegel, H. J. (2015). Adaptive quantum computation in changing environments using projective simulation. *Scientific Reports*, 5, 12874.
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, & W. Donaldson (Eds.), *Organization of Memory* (pp. 2799–2804). Academic Press.
- Van Otterlo, M., & Wiering, M. (2012). Reinforcement learning and Markov decision processes. In M. Van Otterlo, & M. Wiering (Eds.), *Reinforcement learning: State of the art* (pp. 25–52). Springer.
- Wallnöfer, J., Melnikov, A. A., Dür, W., & Briegel, H. J. (2020). Machine learning for long-distance quantum communication. *Physical Review X Quantum*, 1, 010301.
- Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3), 279–292.
- Weimer, H., Müller, M., Lesanovsky, I., Zoller, P., & Büchler, H. P. (2010). A Rydberg quantum simulator. *Nature Physics*, 6(5), 382–388.
- Werner, R. F. (1989). Quantum states with Einstein-Podolsky-Rosen correlations admitting a hidden-variable model. *Physical Review A*, 40, 4277–4281.
- Zwenger, M., Briegel, H. J., & Dür, W. (2016). Measurement-based quantum communication. *Applied Physics B*, 122(3), 50.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 6

Phenomenology



In this chapter, we provide a phenomenological perspective on agency that supports and complements our preceding general treatment of agency. In Chap. 2, we explicated the general problem of agency and contrasted our target notion with the specifically human form of agency that informs most of the philosophical discussion. Certain aspects that are specific to human agency, such as morality, lie outside the notion of agency that we are after. In our analysis and modelling approach, we focused on a notion that is not committed to these aspects, but which is rich enough to capture the agency of (non-human) animals and to allow for the possibility of artificial agency.

While limited in this way, our approach does not exclude all aspects of human agency, or aspects that may appear to be specific to human agency. Notions such as creativity and imagination, as well as the capacity of model-building, are in fact well connected to our general approach to agency. For example, model-building is clearly relevant for scientific agency as discussed in Sects. 4.4 and 5.8.1, and it also offers a perspective on a certain low-level form of agential self-reference.

To prepare the grounds for our phenomenological treatment, we will first motivate why we consider Heideggerian phenomenology to be a suitable and promising framework for understanding a broad notion of agency. While Heidegger's approach originates in a study of the meaning of being and is methodologically tied to specific aspects of *human* agency (*Dasein*), it is in fact also suitable to elucidate general features of agency, and it even provides interesting perspectives on artificial agency. Here we take inspiration from Dreyfus's detailed study of Heidegger and from his phenomenological approach towards understanding the scope and limits of AI.

The central element of the phenomenological analysis that anchors our perspective on agency is Heidegger's notion of *being-in-the-world*. Starting from that notion, we develop the relevant background of the phenomenology of *Dasein*, including the basic idea of encountering entities as *equipment* that is constitutive of being-in-the-world. We draw parallels between equipment and the notion of *affordances* that Gibson develops in his ecological approach to perception. The

unifying claim that we defend is that *agency, in its basic form, is circumspect coping*. Starting in this way, we work towards an interface between the language of existential phenomenology and the language of agency, which we hope is useful also for other researchers.

That approach allows us to discuss what we perceive to be lacunae of the phenomenological approach connected with the basic notions of possibility and of learning, and it provides the basis for a phenomenological reading of Projective Simulation, the agency model discussed in detail in Chap. 5. This in turn offers a fresh perspective on scientific agency—something that is specifically human and yet, probably, a field in which artificial agency will make significant advances in the near future.

6.1 Why Phenomenology?

Philosophical approaches to agency are mostly geared towards the human example, which is usually designated as the full-blooded, ideal case (Ferrero, 2022). Heidegger's phenomenological analysis of *Dasein* in his classic, *Being and Time* (Heidegger, 1927), is no exception—at least at first sight. But while Heidegger indeed targets the adult human being and describes some of the intricacies of human affairs in great detail, his approach mostly avoids the high-level notions of free will and morality that are otherwise typically in focus in philosophical discussions of agency. Even language, propositional thought, and linguistically describable intentions play only a rather minor role in the initial analysis of *Dasein*. That analysis is therefore an interesting and fruitful starting point for a philosophical refocussing on our target notion of agency towards the end of this book. Accordingly, we will use Heidegger's *Being and Time* as the main reference in this Chapter—not from an exegetical point of view, but because of its systematic importance and fruitfulness, which we will try to map out in what follows.

There are in fact several reasons for a phenomenological approach to agency. On the one hand, one can note that the influence of phenomenology is generally rising in cognitive science broadly construed (Gallagher & Zahavi, 2021). On the other hand, with respect to our focus on agency modelling and on the possibility of artificial agency, it is useful to link to Dreyfus's critique of AI, which heavily draws upon Heidegger's phenomenology. In his writings since the early 1970ies, Dreyfus consistently criticised AI as a research program. His criticism was initially aimed at the conceptual framework of symbolic AI and the entire representationalist (Cartesian) way in which AI researchers at the time thought about human intelligence. The field of symbolic AI has indeed failed spectacularly as a research program, at least when measured by the great expectations raised by its leading researchers.¹ It is true that the field of AI has undergone many developments in the last decades,

¹ See, e.g., Minsky (1967, 2).

with impressive successes in specific areas including deep learning and robotics. In our view, however, a convincing model of Artificial General Intelligence should aim at the realisation of artificial agency, which should exhibit basic features of human agency—and for that endeavour, Dreyfus’s main arguments still apply today. At least, it is unclear how current approaches to AI can account for agency (when that term is used in a way that is more specific than the use of “agent” for any system with inputs and outputs that is common in computer science). A phenomenological approach to AI, on the other hand, can put the phenomenon of agency centre stage.

In his criticism of AI, Dreyfus (1992, 2007) was guided by Heidegger’s phenomenology of *Dasein* in *Being and Time* (Heidegger, 1927). Dreyfus (1991, 14) translates the term *Dasein* with the human *way of being*, and for those aspects of *Dasein* that are relevant for his criticism, one can equally use the term *human agency*. This reading of the term *Dasein* as human agency has explicitly been proposed by Wheeler (2005, 122). We will follow this proposal and take being-in-the-world—which is the fundamental structure (*existentiale*) of *Dasein*—as the fundamental structure that underlies human agency. This strong association of being (in-the-world) with acting (in-the-world) conforms to our general explication of agency in Sect. 2.6. We thus transfer the structure of being-in-the-world downwards, as it were, and use it to analyse the agency of animals and, possibly, of artefacts. With this understanding, Heidegger’s phenomenology then also provides a reference point for a theory of artificial agency. Viewed positively, a phenomenological approach can thus connect fruitfully to our discussion of agency. As we have argued in Chap. 2 and illustrated in the context of applications of the model of Projective Simulation in quantum science in Sect. 5.4, our approach to agency also offers novel perspectives on the modelling of artificial agency and AI. We will support this claim by a phenomenological interpretation of PS in Sect. 6.5.

Regarding that point, one may wonder whether a simple model such as PS (or, for that matter, any other existing model of learning) can possibly say anything useful about human agency. In the history of AI, researchers have repeatedly and grossly underestimated the challenge of creating artefacts with human-like intelligence. Nevertheless, formal and mathematically well-specified models can be of great value for defining and then focussing on the essential features of a phenomenon. A historical example is the model of the Turing machine. That model helped Alan Turing not only to formalise the notion of a computation, something that at the time was performed exclusively by humans, but also to investigate the scope of computers and to prove fundamental theorems such as the uncomputability of the *Entscheidungsproblem* (Turing, 1936). Beyond these revolutionary results, Turing’s mechanistic model also offered a way of thinking about (human) computation that had a strong influence on the early field of symbolic and computational AI. A formal model of agency can have similar implications: It not only allows us to ground the philosophical *notion* of agency, but also to understand and develop new perspectives on the *phenomenon* of agency, including the agency of animals. Last, but not least, it may also tell us something about the scope of artificial agency, and provide a foothold to properly judge the rapid developments in the field of artificial intelligence.

6.2 Being-in-the-world as Acting

In the first chapters of *Being and Time*, Heidegger introduces the basic ontological structures (*existentiales*) of *Dasein* and its primary characterisation as *being-in-the-world*.² Of particular relevance for the current purpose is the discussion of the way how *Dasein* encounters other entities in its environment, and the specific kind of awareness and understanding that comes with this.

Being-in-the-world describes the primary situation of every *Dasein*. It is the fundamental situation in which every human being finds itself and with which it has to cope. Every *Dasein* is being-in-the-world. The compound expression is meant to indicate a unified phenomenon. This phenomenon is not composed from independently defined parts, such as “the subject” and “the world”, between which an epistemic relation has to be established. Instead, one can understand in what sense *Dasein* is *in the world*, what *being-in* means, and how one can make sense of the world “as such” by focusing on different structural moments of this unified phenomenon.³

Heidegger’s phenomenological approach is of particular interest for us because it puts being-in-the-world as the basic structure of *human agency* at the centre. It thereby provides a reference also for the phenomenological interpretation of a more general notion of agency that we target in this book and which includes the agency of animals and, possibly, future AI artefacts. Consciousness or conscious mental states play no explicit role in the existential analysis of *Dasein*—*Dasein* is not a conscious subject (Dreyfus, 1991, 13). This accords well with the minimal role that mentalistic terms play in our explication of agency in Chap. 2, which avoids the notion of consciousness. A further parallel lies in the limited role of the notion of intentionality, for which we suggested the less loaded terminology of “wanting” in Sect. 2.6.1.iii. Most importantly, neither Heidegger nor our approach to agency is committed to action explanation via propositionally structured intentions, beliefs, and desires as independently given mental states.⁴

In the following section, we give a brief summary of the phenomenological framework, focusing on those parts that seem relevant for the problem of agency in the general sense.

² See Part 1, in particular Chapters 2, 3, and 5 of Heidegger (1927).

³ Heidegger’s analysis has its own methodology, existential-ontological hermeneutics, which introduces fundamental structures called *existentiales* that are seen as constitutive for *Dasein*. They are to be distinguished from so-called categorical structures that are normally used for the characterisation of entities different from *Dasein* (e.g., how matter and things are composed, their elementary constituents, properties, and interactions). We will not discuss Heidegger’s methodology further here.

⁴ Such action explanations are typical of a causal approach to agency in the tradition of Davidson (1963). Searle (1983) also broadly belongs to this tradition. With reference to Searle, but also reaching further back, Dreyfus (1993) points out crucial differences between Heidegger and the Husserlean tradition when it comes to intentionality.

6.2.1 *Being Absorbed: Equipment, Involvement and Reference, Circumspection, and Coping*

One of the first targets of the analysis of being-in-the-world in Heidegger's *Being and Time* is the environment and the ontological status of entities encountered by *Dasein* in its everydayness.⁵ The everyday way of encountering entities is not a contemplative and detached one, but an active, *absorbed* and *concerned* way of dealing with entities, which can also be characterised as absorbed and concerned *coping* (Dreyfus, 1991). Heidegger calls entities encountered in this way *Zeug*, translated into English as *equipment* (Heidegger, 1962; Dreyfus, 1991) or *useful things* (Heidegger, 1996).⁶

This absorbed and concerned way of *Dasein*'s dealing with entities is not understood as a conscious and intentional process in which a subject interacts with objects in an otherwise independently given and determined world. The ontological status of equipment is different from that of merely occurrent things that have a prior and independent categorical characterisation. Equipment has its meaning only relative to *Dasein*, that is, in the context of (human) agency. The primary characterisation of equipment is: something *in-order-to*. Equipment acquires its meaning from what an agent can possibly *do* with it. The hammer, one of Heidegger's favourite examples for illustration, is something to be used *in order to* drive a nail into the wall, work on malleable iron, or break a window *in order to* get out of a locked room when the house is on fire, etc. To emphasise the agent-relative status of equipment, as something whose meaning is given only in the context of agency, Heidegger calls this status *ready-to-hand*, contrasting it with the status of entities that "simply occur", or are *present-at-hand*, when viewed from a detached, third-person perspective. The being of equipment is thus that it is *ready-to-hand*, to be used *for-the-sake-of* something, as in the example of the hammer that is used to break the window *for-the-sake-of* escaping from the fire.

In this phenomenological reading, the environment of the agent thus consists of, or, better, is encountered through, equipment. Since the meaning of equipment, its *in-order-to*, only appears in a larger context, there is, strictly speaking, nothing like a *single*, isolated piece of equipment. Equipmental entities always refer to other equipmental entities, like the hammer referring to the window, which has its own *in-order-to*, namely to protect the agent from the cold weather outside, etc. Every piece of equipment refers to something else, and this referring-to-something-else is part of

⁵ See the discussion in Chapter 3, A, Section 15 of Heidegger (1927).

⁶ The translation of *Zeug* as *equipment* (Macquarrie and Robertson; Heidegger, 1962) is, in fact, too rigid and suggests something prefabricated, whose purpose is already determined, thus depriving the agent of some of its initiative. This becomes especially visible and relevant when we consider *Zeug* and actions in the context of child agency. The translation as *useful things* (Stambaugh; Heidegger, 1996) seems better, but it has the disadvantage that it consists of two words in which *useful* appears as an attribute of *thing*, which suggests a primary characterisation of *Zeug* as an occurring thing. Heidegger (1927, 96 [68]) refers to *πράγματα* as a fitting Greek expression.

a *totality of reference* (*Verweisungszusammenhang*) and *involvement* (*Bewandtnis*). This background of referential structures underlies the agent's basic grasp of its environment, which Heidegger calls *circumspection* (*Umsicht*). According to Dreyfus (1991, 68), circumspection can be seen as a mode of awareness, before and outside any theory of cognition. Dreyfus (1991, 67) elaborates on the notion of awareness as follows:

According to this philosophically unprejudiced description of everyday skillful coping, there is awareness but no self-awareness. That is, there is no self-referential experience of acting as this is understood by Searle (and would have been understood by Husserl), i.e., no experience of volition with the conditions of satisfaction that this experience of acting cause the action. As Heidegger puts it: "Self and world belong together in the single entity, Dasein. Self and world are not two entities, like subject and object [...] but self and world are the basic determination of Dasein itself in the unity of the structure of being-in-the-world" (Heidegger, 1982, 297). Or, even more directly, "Dasein [...] is nothing but [...] concerned absorption in the world." (Heidegger, 1985, 197)

Similarly, reference and involvement also give rise to the notion of *significance* (*Bedeutsamkeit*), which is "the background upon which entities can make sense and activities can have a point" (Dreyfus, 1991, 97).

6.2.2 A Graph-theoretical Analogy

The language of phenomenological analysis and the various terms used to describe the way of encountering entities may be unfamiliar. For those with a more formal background, we offer the following graph-theoretical analogy, which may help to associate some of the basic features of these terms and to keep them apart:

- A natural picture represents the totality of reference—the referential manifold—as a graph, where the instances of *equipment* (*Zeug*) are nodes (also called vertices) and *references* are directed edges connecting different nodes. As an example, Fig. 6.1 illustrates parts of the referential manifold of a hammer. If one imagines that each edge carries the label "in-order-to", that figure naturally encodes the example of the hammer that we gave above.
- The notion of *involvement* refers to other, more indirect aspects of equipment. In our graphical illustration, involvement corresponds to certain non-local aspects of the graph, beyond the direct neighbourhood of a certain vertex. For example, in the practical context described in Fig. 6.1 building a hut is a possible goal towards which the activity of hammering is directed, while getting shelter is one of its possible ends, for the sake of which the agent is engaged in her activity.⁷ Involvement can thus be thought of as highlighting various (end)points of the

⁷ Dreyfus (1991) distinguishes a "where-in" (or practical context), a "with-which" (or item of equipment), a "towards-which" (or goal), and a "for-the-sake-of-which" (or final point), to describe how our everyday activities like hammering or writing on a blackboard make sense. He illustrates this with the following example: "I write on the blackboard *in* a classroom, *with* a piece of chalk,

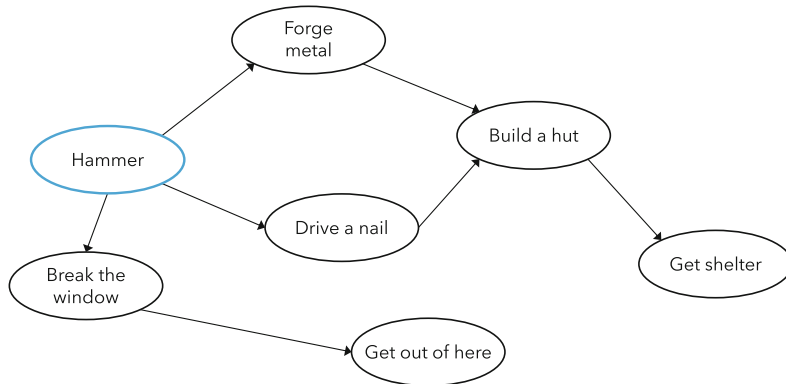


Fig. 6.1 Graphical representation illustrating part of the referential manifold of the hammer as equipment

graph where paths from a given vertex can lead to, possibly together with other vertices that are encountered on these paths.

- The notion of *significance*, in turn, highlights the graphical structure (referential manifold) *as a whole* that surrounds a specific vertex (equipment). This captures the various ways in which equipment is connected and offers paths towards different ends and purposes.

Of course, the illustration of ontological vocabulary in terms of a simple graph needs to be taken with proper care and reservation. Indeed, Heidegger warns against flattening the idea of a referential context as a mere relational system:

The context of [...] references, which, as significance, is constitutive for worldhood, can be taken formally in the sense of a system of Relations. But one must note that in such formalizations the phenomena get levelled off so much that their real phenomenal content may be lost, especially in the case of such ‘simple’ relationships as those which lurk in significance. (Heidegger, 1927, 121 [88])

One should thus understand the proper value of a graphical illustration as in Fig. 6.1, which is not meant to suggest that a directed arrow in that figure is all there is to the ontological notion of reference as introduced by Heidegger. Primarily, it just means that the characterisation of the hammer that we wrote down in a few lines of text can equally well be represented by that graph. At the same time, however, that graph can also be a starting point for more operational formalisations. Such formalisations may then serve as methodological vehicles into areas where there is no longer an experiential foothold for the phenomenological method. These areas include the agency of non-human animals and of artefacts. Indeed, in Sect. 6.5 we will interpret the referential manifold of equipment in terms of the memory structure of a PS agent

in order to draw a chart, as a step towards explaining Heidegger, for the sake of my being a good teacher” (Dreyfus, 1991, 92).

as described in Sect. 5.2, the Episodic and Compositional Memory (ECM), which has an explicit graph structure (see Figs. 6.2 and 6.3).

6.2.3 *Equipment and Affordances*

Reference and involvement cannot be found in the things “in themselves”, but they are constitutive of *Dasein*’s way of encountering entities. Equipment is equipment only for *Dasein*, only relative to an agent, and in the referential context thus always lies an individual agent’s perspective on the world. For the one who takes care of the fire, a stone as equipment may refer to the possible containment of the fireplace. For the other, it may refer to hunting as a throwing tool. Similarly, a hammer may refer to all kinds of purposes, but only for the one who knows how to use it, either because it was shown to her or because she found a new application herself. For the uninitiated, the hammer on the first encounter is perhaps a mere piece of metal with a wooden handle. To know what a hammer *is*, one needs to know what it is *good for*, and in this case this includes knowing what it was *made for* (by others). The same is true for any kind of tool and to an even greater extent for scientific instruments.

Above we already mentioned circumspection as the agent’s basic mode of awareness. Circumspection means that the agent in her acting—that is, in the simplest form, in her dealing with equipment—always *sees* different possibilities. This seeing is not a thematic-categorical grasp of a domain of objects. Rather, it denotes a space of references that comes with any given perceptual situation and opens up *for* the agent. The stone on the trail may refer to the possibility of stumbling over it, or of avoiding it, or it may refer to the possibility of using it as a throwing tool. This referring to possibilities, however, does not originate from and is not a property of the stone considered in itself. It is connected both to the individual (learning) history of the human agent in her dealing with equipment—in this case, with stones—and to the evolutionary history of the human species, including the development of the human hand and its suitability for grasping and throwing.

This evolutionary interpretation connects nicely with the notion of *affordances* in Gibson’s (1986) theory of perception. Affordances characterise entities in the environment of the animal which provide *possibilities for* the animal:

The *affordances* of the environment are what it *offers* the animal, what it *provides* or *furnishes*, either for good or ill. The verb *to afford* is found in the dictionary, the noun *affordance* is not. I have made it up. I mean by it something that refers to both the environment and the animal in a way that no existing term does. It implies the complementarity of the animal and the environment. (Gibson, 1986, 127)

Affordances, in the theory of animal perception, thus have a similar status as equipment for the human agent. Affordances can be understood as referential structure *for the species* that has developed under evolution and which has been mapped into the genetic level, or as referential structure for the individual animal agent that it may have gained through its individual or social learning history.

As we already said, reference and involvement are parts of the unique phenomenon of being-in-the-world that characterises (human) agency. The possible ends of a specific equipment-entity in the agent's environment (i.e., the entity's involvement) is neither just conceived by the agent, nor is it a property that can be attributed to the equipment-entity alone. It is rather a feature of the human way of *being-in* and thus lies, as it were, in-between the agent and the environment. This in-between-ness, between the agent, on the one hand, and her environment, on the other, is what the phenomenological notion of equipment shares with Gibson's notion of affordances:

An important fact about the affordances of the environment is that they are in a sense objective, real, and physical, unlike values and meanings, which are often supposed to be subjective, phenomenal, and mental. But, actually, an affordance is neither an objective property nor a subjective property; or it is both if you like. An affordance cuts across the dichotomy of subjective-objective and helps us to understand its inadequacy. It is equally a fact of the environment and a fact of behavior. It is both physical and psychical, yet neither. An affordance points both ways, to the environment and to the observer. (Gibson, 1986, 129)

For our approach to agency, the most important similarity between affordances and equipment is this in-between-ness, their pointing both to the environment and to the agent. There are further similarities but also differences. The typical example of equipment, the hammer used in the workshop, is an artefact made by humans for other humans, and its use is learned by the individual *Dasein* in a social setting. The typical example of an affordance, the flat and stable ground that a heavy terrestrial animal can stand and walk on, is a naturally occurring feature of the environment that is reflected in the bodily structure and thus in the genetic make-up of the species, and its use is, for most species, an innate capability. These examples are thus strongly opposed. But there are also convergences between equipment and affordances—humans also rely on given environmental features that are reflected in their bodily and genetic make-up and whose use comes naturally to the individual, and some non-human animals can also learn to use tools individually and socially, and many species also build artefacts such as nests or dams.

An interesting question that arises at this point is which notion of a *world* becomes available through these different agents' dealings with their environment via equipment and affordances. This connects to the phenomenon of *breakdown*, which we have not yet discussed.⁸ Breakdown is a phenomenological term that describes a situation when equipment malfunctions or is unavailable, breaking the flow of smooth coping and thus drawing the agent's attention to the equipment's various other references. Making these (previously non-thematic) references manifest then corresponds to a mode of encountering that reveals the availability or unavailability of equipment in a thematic way. Breakdown thus plays a key role in revealing the *worldliness* of environmental entities and their relations, thereby constituting the first step towards an explicit idea of *the world*.

⁸ For a useful overview with references to Heidegger, see Dreyfus (1991, 70ff).

Breakdown as a world-disclosing phenomenon seems to make sense only for the individual agent. If one were to insist on a strict separation between affordances, as referential structure just for the species, and equipment, as referential structure just for the individual human agent, then it would follow that the individual animal has no means of accessing that structure, and that the notion of breakdown would not apply to non-human animal agents. These could therefore not disclose a world via breakdown either. Heidegger's own analysis may point in this direction, especially given his use of artefacts as instances of equipment,⁹ but such a strict separation is in fact unwarranted and unnecessary. In Sect. 6.4 we will pick up this discussion in the context of learning, where the ontological status of equipment and its referential structure will be reconsidered.

6.2.3.1 Comparing Gibson and Heidegger

Gibson and Heidegger come from different directions. Gibson's notion of affordances is part of a psychological-ecological theory of perception that focuses on non-human animals. Heidegger's phenomenological analysis of *Dasein* is part of a more fundamental approach towards the understanding of the meaning of being. His analysis of *Dasein* in its everydayness deals with the specific *human* way of being. But while the meanings of affordances and equipment are not the same from the outset, if we expand the notion of equipment beyond the human examples and broaden Heidegger's focus on a structure of artefacts, their content becomes almost identical. Their main characteristic is their pointing both ways, to the agent and to the environment. It is worth mentioning that Gibson's theory of perception and the notion of affordances have come to play an important role in robotics (Horton et al., 2012; Bohg et al., 2017), while Heidegger's analysis of *Dasein* has influenced developments in artificial intelligence towards *embodied AI*, notably through the work of Winograd and Flores (1987), Dreyfus (1992), and others.¹⁰ So here, too, there seems to be a strong connection.

Circumspection as agential awareness captures the kind of "seeing" of possibilities that equipment and affordances offer to an agent. It connects with other *existentiales*, including *understanding* and *projection*, which will be discussed in the following Sects. 6.3 and 6.5. For our analysis of agency, a key aspect of circumspection is its action-guiding role that enables an agent to cope with its environment. Dreyfus (1991, 66f.) describes circumspection as a non-thematic form of awareness which is best exhibited when an agent is fully absorbed in its activity, coping smoothly and skilfully. Circumspection can be interpreted as a continuous

⁹ Heidegger himself makes this separation explicit via his thesis that "the animal is poor in world" (Heidegger, 1995, 185ff.).

¹⁰ Further developments in that discussion are provided by Norman's interpretation of affordances in the context of design (Norman, 2013), including, e.g., the notion of perceived affordances, which has found widespread applications in the field of human-computer interaction. For a recent review see Ardón et al. (2020); a recent application is described by Bahl et al. (2023).

projection into possibilities. For *Dasein* as a historical and individual being, the scope of circumspection and projection can change over time. In other words: what is seen by circumspection, the uncovered referential structure that constitutes equipment, changes in the course of learning. Such a diachronic interpretation of circumspection leaves the ontological status of equipment intact. This may change if we interpret learning more broadly, beyond the mere acquisition and successive uncovering of a given context of references. Learning in this broader sense also includes the process of merely playing around with equipment, thereby creating completely new referential contexts and possibly entirely new uses of equipment (see Sect. 6.4.2).

6.2.3.2 Agency as Circumspect Coping

One can summarise our discussion so far via the thesis that, from a phenomenological perspective, *agency, in its basic form, is circumspect coping*. This can serve as a starting point to draw further connections between the structures of phenomenology and agency. However, there are at least two challenges if one tries to defend this thesis. These challenges have to do with the historicity and individuality of agents as *settlers* and *learners* (see Sect. 2.6). Insofar as learning is constitutive for agency—already in its basic form—it must be accountable for within in the phenomenological-existential framework. Circumspect coping must thus come with a built-in notion of flexibility and adaptivity. Flexibility could be attributed to the way in which *Dasein* (or, more broadly, an agent) may explore and exploit the referential manifold that underlies circumspection. Adaptivity could be attributed to the way in which the referential manifold itself (which is constitutive of and defines equipment) may change over time.

The first point, flexibility, raises the challenge to find a place for the indeterminism that is necessary for agency in the framework of phenomenology. The second point, adaptivity and learning, raises the challenge to ultimately reconsider the ontological status of equipment. These two challenges will be discussed in more detail in the following Sects. 6.3 and 6.4. As we will see, these challenges also provide opportunities to establish fruitful connections between phenomenology and quantum physics as discussed in Chaps. 3 and 4.

6.3 Indeterminism and Existential Possibilities

We have seen that being-in-the-world as acting in the sense of circumspect coping means that an agent operates with a certain room to manoeuvre, within a specific space of possibilities. In Heidegger's own description centring on human *Dasein*, this room to manoeuvre is primarily given through the referential structure of equipment that the agent deals with; in Gibson's description of animal perception and action, the room to manoeuvre is given through the affordances of the environment. In both cases, we consider not just any old open possibilities, but

possibilities that make sense for the agent. And we consider a way of dealing with these possibilities that is not characterised by the distanced evaluation of available options and the selection (or perhaps even calculation) of the optimal next move, but by the agent's direct and smooth handling of a given situation.

In view of our focus on agency modelling, it is important to characterise the relevant possibilities in a way that is both scientifically tractable and true to the spirit of the phenomenological approach. We claim that the kind of possibilities that is at issue here can indeed be elucidated by starting from Heidegger's own description, once we have situated it before the background of our analysis of agency and indeterminism from Chap. 2. We want to stress again that we do not intend our discussion to be in line with any existing school of phenomenological philosophy, nor to be exegetical. Our point is, rather, to try to draw out interesting conceptual connections that should be useful for philosophers, physicists, and AI researchers alike. As in most other parts of this Chapter, we therefore proceed by translating the structures identified through Heidegger's phenomenological analysis of human *Dasein* down to the level of agency in general.¹¹ Our focus is therefore on agential possibilities and their role for being-in-the-world, i.e., in circumspect coping through dealing with equipment—but at a level of complexity below full-blown *Dasein*. This will ultimately put us in a position to indicate how other parts of the phenomenological analysis, notably the notions of thrownness, understanding, and projection, also find their natural place in the resulting picture.

It will be good, anyway, to start from a concrete textual source. In *Being and Time*, Heidegger emphasises that the existential possibilities involved in *Dasein*'s coping must be distinguished from other, more basic kinds of possibility, because *Dasein* is *existentially tied* to possibilities. It is worth quoting and commenting on the relevant passage here:

Dasein is not something present-at-hand which possesses its competence for something by way of an extra; it is primarily Being-possible. Dasein is in every case what it can be, and in the way in which it is its possibility. [...] The Being-possible which Dasein is existentially in every case, is to be sharply distinguished both from empty logical possibility and from the contingency of something present-at-hand, so far as with the present-at-hand this or that can 'come to pass'. As a modal category of presence-at-hand, possibility signifies what is *not yet* actual and what is *not at any time* necessary. It characterises the *merely* possible. Ontologically it is on a lower level than actuality and necessity. On the other hand, possibility as an *existentiale* is the most primordial and ultimate positive way in which Dasein is characterised ontologically. [...] Possibility, as an *existentiale*, does not signify a free-floating potentiality-for-Being in the sense of the 'liberty of indifference' (*libertas indifferentiae*). [...] Dasein is Being-possible which has been delivered over to itself—*thrown possibility* through and through. Dasein is the possibility of Being-free for its ownmost potentiality-for-Being. (Heidegger, 1927, 183 [143f.])

We will argue that we can connect the points that Heidegger makes here with our explication of agency from Chap. 2 and especially with our discussion of

¹¹ In analogy with Thompson (2004), we can speak of different “conceptual gears” within which the relevant agential structures can be identified.

the ontological emergence of possibilities based on fundamental indeterminism (Sect. 2.2.3.4).

6.3.1 *Heidegger on Existential Possibilities*

In the above quote, Heidegger points to the specific, existential notion of possibility that characterises *Dasein* as being-possible. Focusing on the agency aspect of *Dasein*, this translates into the claim that agents are fundamentally characterised via their possibilities—these do not come “by way of an extra” added to a thing that is otherwise already fully formed. We agree: agents must be settlers of open possibilities and thus cannot be determined; they would cease to be if no possibilities were open for them (see Sect. 2.6.1.ii). Adapting the Aristotelian dictum that for a living thing, to be is to live, we can say that *for an agent, to be is to have meaningful open possibilities*. In Sect. 2.2.3 we have defended the more general claim that no higher order of organisation can be real at all unless it exists on the basis of underlying open possibilities of its matter that can be organised as material. For agency, as pointed out in Sect. 2.2.3.4, we have to presuppose a multi-levelled stacking of such levels of organisation: The agent’s body is organised, on the basis of its material (e.g., the living tissue) that supplies the necessary open possibilities, but that organised body still leaves open further, higher-level possibilities that *the agent* can organise through its various actions. Thus, we can translate the possibility as *existentiale* that Heidegger finds in his characterisation of the being of *Dasein* as the higher-order meaningful possibilities that any agent needs to have available in acting. Abstracting from the specifically human case, we can also connect these possibilities with Gibsonian affordances, as already mentioned. Without such possibilities, without the capacity to meaningfully settle matters pertaining to the movement of their bodies, agents—whether human, animal, or artificial—cannot exist.

Heidegger contrasts possibility as an *existentiale*, which characterises *Dasein*, with a notion of “empty logical possibility” that he rejects. That latter notion is linked to the traditional (Kantian) distinction of three mutually exclusive modal categories of actuality, necessity, and mere possibility.¹² Possibility in this purely logically determined sense characterises that which is neither actual (realised) nor necessary (unavoidable), i.e., that which as yet both may or may not be. As mere contingency, this kind of possibility is, as Heidegger says, ontologically “on a lower level than actuality and necessity”. He describes the distinction of this kind of possibility from actuality and necessity in temporal terms (“*not yet*”, “*not at any time*”). It seems appropriate for our discussion, then, to equate (mere) possibility in the purely logical sense that Heidegger rejects in the above quote with an abstract

¹² In this sense, possibility occurs as one of the concepts of pure understanding in Kant’s table of the categories (Kant, 1781, A80/B106), where it belongs to the class of modality.

notion of unactualised possibilities for the future. In the context of our discussion of ontological emergence we have seen that possibilities for the future may be open with respect to different levels of organisation, and that open possibilities at lower levels may be closed off at a higher level.¹³ At the most basic level, such indeterministic possibilities are grounded in the fundamental indeterminism of quantum mechanics, as discussed in Sect. 3.5. The abstract, merely logical notion of possibility that Heidegger rejects as irrelevant for existential possibility is thus already exemplified at the lowest ontological level of quantum indeterminism. We can then understand Heidegger as saying that such base-level indeterminism does not (yet) provide an adequate picture of agential possibilities. We agree.

In describing ontological emergence, however, we stressed that base-level indeterminism is a necessary, but by no means a sufficient condition for the emergence of higher-level structures, including higher-level possibilities such as the possibilities for an agent to *move its body*. In discussing agential possibilities, we pointed out that the open possibilities that underwrite an agent's non-rigidity (non-determinedness) have to fulfil further, agent-specific conditions. First, temporally, they have to span different time scales to enable short-term flexibility and long-term learning and adaptation. Second, qua content, they have to be meaningful for the agent.

As we pointed out in Chap. 2, most of the agency literature ignores the longer-term perspective on agency that is central to our approach. Heidegger's discussion of *Dasein* is no exception—his focus is on fully developed, competent being-in-the-world, not on the development of agency. As we will further discuss with specific reference to phenomenology in Sect. 6.4 below, however, learning and adaptation are crucially important for modelling agency and for understanding how agency can be real—how agency can be understood as a phenomenon in our natural world. The different temporal scales of an evolutionary perspective and of a focus on an individual agent's history can help to establish a more detailed account of how lower-level possibilities are organised into meaningful agential possibilities. If such a perspective is lacking and the category of possibility is understood in a narrow sense that lumps together all forms of indeterminism irrespective of their level of organisation, the only description of open possibilities that are formally an agent's is in terms of a liberty of indifference. Heidegger is right to reject this reading of existential possibilities as too meagre; a liberty of indifference does not properly characterise agential possibilities. A merely “free-floating potentiality-for-Being” does not yet anchor meaningful possibilities. Rather, it gives rise to a wrong conception of indeterminism-based free agency as a matter of sheer luck. Agency modelling must provide an account of *sensible*, meaningful actions or activities, it cannot be limited to open possibilities simply turning out one way or the other. A diachronic perspective is helpful at this point.

¹³ See, for example, our discussion of bacterial chemotaxis in Sect. 2.4.1.2.

6.3.2 *Thrownness and the Fit of Agential Possibilities with the Environment*

The required long-term perspective helps anchor a proper notion of sense by showing how *agential open possibilities are non-accidental*. Consider Gibsonian affordances first. The animal evolves in an environment that, to first approximation, defines a given niche into which the animal has to fit.¹⁴ The natural conditions of that environment (the landscape, the vegetation, the abundance of water and shelter, the prevalence of other animals, some of which may be predators or possible prey) offer structure that—over a very long time—an animal species can integrate into its bodily and behavioural organisation, thus giving rise to an environment that is structured in terms of affordances. Given such a background history, it is thus not an accident that the animal finds itself in an environment offering meaningful possibilities—if this were not so, the animal species would not have survived. Life is proof of adaptation. For the higher-level, existential possibilities that Heidegger describes in his analysis of *Dasein*, a similar story can be told (even though, as we will discuss in Sect. 6.4, we need to add a developmental side that is lacking in his analysis): equipment is shaped in a referential nexus anchored in (social) practices, e.g., of a trade, and therefore it is no accident that equipment affords circumspect coping. Biological agency, whether at the level of non-human animals or at the level of *Dasein*, can thus be expected to be situated in an environment that offers meaningful possibilities for action as the agent's way of being-in-the-world.

This does not mean, of course, that things cannot go wrong. We mentioned the phenomenon of breakdown in Sect. 6.2.3. Agents can find themselves in situations that obstruct their agency, and a situation may even fail to afford any meaningful ways of smoothly going about. No agent is in full control of the situation it finds itself in—fundamentally, an agent is always *thrown into* a situation. For Heidegger's analysis of *Dasein*, this thrownness is a crucial aspect of existence. Many features of thrownness are specific to the human case, and we will leave them by the wayside here. For example, in Heidegger's analysis it is crucial that *Dasein* is concerned with (*cares for*) its own being, always in danger of losing sight of this concern, and that it has to come to grips with the fact that its history will terminate in death. Such high-level considerations will be discussed to some extent in Sect. 6.6.1. They are not necessarily present in all agents. It is true that as material beings, all agents can persist for just a limited period of time. But agents need not be aware of this fact; indeed, such awareness presupposes a fairly sophisticated type of self-thematisation that may only be possible for linguistic beings. Agents, as we already pointed out in Sect. 2.6, need not be self-aware in this sense, and need not have a language. The notion of thrownness, however, still makes sense also for less sophisticated agents: all agents are at best in limited control of their situation.

¹⁴ Some animals—well-known examples being leaf-cutter ants, beavers, and us humans—also considerably shape their environment to fit *them*, but this is best understood as a refinement of a basic process of fitting in, which remains our focus.

The developmental argument for non-accidentality sketched above indicates that natural agents thrown into a situation normally find themselves confronted with (at least some) meaningful possibilities, so that the default mode of being-in-the-world as circumspect coping can normally take its course. In case such coping is frustrated and breaks down, agents will have to back up, as it were, and reconsider. We already mentioned that Heidegger describes how *Dasein* in such a case of breakdown can switch from an engaged and absorbed, non-thematic dealing with equipment to a thematic, distanced consideration of, e.g., broken or unsuitable equipment—the nail that is bent, the hammer that is too heavy, the piece of wood that splinters. Whether other, non-human agents will be able to switch mode in such cases depends on the sophistication of their organisation. In describing agents fundamentally as learning entities whose behaviour is not fixed in a rigid way, we affirm that some short- and long-term flexibility and adaptability must always be present. It is also true, however, that for any given agent there are environments that are simply too challenging. To give an extreme example, no agent can cope with being thrown into a black hole.

For artificial agents, the mentioned developmental argument is not necessarily available. Artificial agents can be constituted in such a way that they are unsuited for the environment that their makers want to put them in. If that happens, artificial agency can be frustrated from the start, like the development of an acorn falling onto a concrete pavement. Without a natural history shaping the agent to fit into its environment, there is no internal connection between an artificial agent's capabilities and the environment in which it will be deployed—the connection is, to begin with, merely accidental. It is up to the designer of the artificial agent to guarantee a good fit. In limited, tightly constrained environments, such a fit can perhaps be engineered from the start, but in most cases, learning and adaptation will be crucial for success.¹⁵ This implies that the thrownness of artificial agents is more external to their constitution and their capabilities than for natural agents. The notion, however, still applies.

6.3.3 *Projection and Understanding*

Thrownness, transposed down from *Dasein* to the level of mere agency, means that the agent has to come to grips with its environment, that it has to make sense of it, that it has to *understand* it—not in the sense of having a fitting representation of the environment, but in the sense of fitting in the environment itself. Such understanding shows itself in the agent's smooth coping. Heidegger describes this understanding in terms of the agent's *projection* (*Entwurf*).¹⁶ Projection is a non-

¹⁵ See Braitenberg (1984) for an interesting early description of (thought-)experiments pointing that way.

¹⁶ Heidegger's German terminology exposes the tight connection between thrownness (*Geworfenheit*) and projection (*Entwurf*) through the common root "werfen", to throw. There appears to be no good corresponding English translation. The common root, however, occurs in "projection" at least in its Latin variant ("iacere", to throw).

thematic way of sampling from the available, meaningful agential open possibilities. In their circumspect, absorbed coping, agents always project themselves into the future. This process is a core part of the agent's agency, and therefore it is never finished—the agent is always, as it were, ahead of itself in its projection, and will be confronted with the basic facticity of what happens. Agency is always directed towards the fundamentally open future, and the agent expresses its non-thematic understanding of its situation and its meaningful options via the absorbed coping with that situation.

In Heidegger's analysis of human *Dasein*, thrownness also has a strongly existentialist connotation that is intimately connected to the fact that *Dasein* necessarily *cares* for its being: *Dasein* is always also thrown into the care for its own being. This self-reflective structure need not be present in all agents, as we just pointed out; in fact, in its full form it appears to be a specifically human trait, and we will not discuss it here.¹⁷ We thus also leave by the wayside the specifically human ways of failure in the proper care for one's being that Heidegger analyses in great detail in terms of authenticity. An agent need not have the conceptual resources for conceiving of itself as authentic vs. inauthentic or alienated. The basic structure that in the case of *Dasein* grounds the polarity of authenticity vs. inauthenticity is, however, present in agency as such: agents are primarily *absorbed* in their coping with their environment, pressing into meaningful open possibilities by way of projection.

We are thus in agreement with Heidegger as regards the critical points he raises against notions of possibility that must be rejected as too meagre, as not living up to the demands of explicating possibility as an *existentiale*. Heidegger (1927, 183 [143f.]) says that existential possibility provides “the most primordial and ultimate positive way in which *Dasein* is characterised ontologically”. This agrees with our slogan that for an agent, to be is to have meaningful open possibilities. Heidegger's own characterisation of existential possibility, however, remains negative. Thrownness and the facticity of the agent's situation provide an anchor for the agent's absorbed coping, but there is no account of the *fit* of agent and environment that makes that kind of coping possible. Above we have already indicated that for natural agents, their evolutionary history in terms of the emergence of higher-level ontological possibilities provides a positive account of that fit—of the fact that the agent's environment affords meaningful possibilities for it. (Bluntly speaking, otherwise the agent would not be there.) From the point of view of agency modelling, that fit is, however, more of a challenge, and the notion of learning and adaptation needs to be considered in detail. Such a discussion is also a lacuna of Heidegger's analysis of *Dasein*.

¹⁷ We will later discuss a minimal form of self-reference that comes with the projective structure of an agent's being ahead of itself; see Sect. 6.6.1.

Even with respect to the case of human *Dasein*, that lacuna has been pointed out—witness Dreyfus, who tries to give a response on behalf of Heidegger:

One might reasonably ask, how can *Dasein* have *always already* projected? Must it not begin sometime? By now, however, it should be clear what Heidegger's answer to this question would be. Of course the human organism must at some time begin to take a stand on itself by pressing into human possibilities. It cannot do this just by reflex action or even by animal directedness. Before it can humanly cope, the baby must be socialized into shared, ongoing activities by imitating people and accumulating the necessary experiences until it begins to do what one does for-the-sake-of whatever it is one is. As soon as the baby is seen as up to something, i.e., its activity can be seen as making sense, then it can be seen as *Daseining*, i.e., as *already* projecting on possibilities. We have seen the same argument, that *Dasein* is *always already* X-ing, applied to moods and to being with others, and we shall see it again when we come to falling and telling. Of course, the whole *Dasein* structure does not suddenly take over the baby. The organism starts *Daseining* gradually. As Wittgenstein says in a similar context concerning language: "Light dawns gradually over the whole."¹⁸ But to be said to have *Dasein* in it, the organism must already be pressing into social possibilities that matter to it. (Dreyfus, 1991, 187f.)

More broadly speaking, thus, we again see that learning is absolutely basic for agency.

6.4 Learning

As we have argued at the end of Sect. 6.2.3, the phenomenological interpretation of agency requires a more explicit discussion of learning in terms of the existential structures of *Dasein*.

Consider a simple example of a human action: driving a nail into a board with the help of a hammer.¹⁹ The skilful execution of this action needs to be learned. We might try to learn it by watching someone else, for example the skilled carpenter, at work and then come up with a detailed description of how she does it: where and at what angle she holds the nail, how far she swings the hammer, with what force she strikes the nail, and so on. Such a description may be provided by the carpenter herself, when she explains to us how she does it, what to pay attention to, etc. Typically such an explanation, or instruction, will come together with an actual demonstration. At some point, we may also be able to give an explanation that is based on our own experience, once we have successfully learned how to use the hammer. What is relevant in any of these cases is that the successful hammering will be the result of an active learning process, during which we will also miss the nail once in a while and perhaps hit the finger, or the nail may fly away or get stuck crookedly in the wood. During the learning phase, in which the correct use of the hammer is learned and practised, our execution of the action will be different from

¹⁸ That quote is from Wittgenstein's last work, *On Certainty* (Wittgenstein, 1969, §141).

¹⁹ See also Figs. 6.2 and 6.3 in the following Sect. 6.5, where we will revisit this example in the context of the PS model.

that of the skilled craftswoman; for example, we may pay more explicit attention to holding the nail correctly, be perhaps tense and insecure, and still have fresh memories of painful failures or the mockery of the others. All of these diverse sensory and motor impressions and other memories are part of the learning process; they accompany us for some time and remain part of the prudent use of the hammer, for each of us in our own way.

The type of learning that is described here is *learning how-to*. In the current context, it is learning *how to hammer* or, more precisely, how to use the hammer in order to drive a nail into a board. This is an example of learning to *skilfully cope* with a given task or situation. In the above description one can also see that the learning process comes with a notion of flexibility and adaptivity. Often something will go wrong, but over time we will improve, hopefully with the help of a teacher's instructions on what to pay attention to, how to make the nail go smoothly into the wood, not bend the nail, not hit the fingers, etc. Adaptivity refers to a time scale over which we perform a similar task several times, thereby hopefully becoming more and more skilful and confident in striking, and able to cope with different sorts of hammers and nails. Flexibility refers to the shorter time scale of each concrete individual action: we may not always hold the hammer in the same place on the handle, and not always apply it at the same angle, etc., but still be successful in driving the nail into the wood. Flexibility and adaptivity also characterise the variability of the experiences that make up the learning-historical background of an action and that shape the agent's memory.

The ontological structure behind skilful coping is circumspection. To understand how skilful coping can be learned within the existential-ontological framework, we need to understand how an individual agent's circumspection can change as the agent learns and practises. A term used by Heidegger that relates to learning is *uncovering*. In the specific context of hammering, he writes

The hammering itself uncovers the specific 'manipulability' ["Handlichkeit"] of the hammer. (Heidegger, 1927, 98 [69])

Here, uncovering should not be understood in an epistemic sense (exhibiting what was already there), but as a way of actively revealing the involvements of equipment. A similar example would be the uncovering of the different visual appearances and the accessibility of an entity through bodily movements of the agent, a topic that has also been discussed under the term *interactive perception* in robotics (Bohg et al., 2017). Another example would be the uncovering of certain haptic sensations or of the handiness of an entity when touching its surface. Hammering will uncover the accessibility and handiness of the hammer including other sensori-motor impressions that come with the more or less skilful use of the hammer.

Learning of this type will be further investigated in Sect. 6.5, where we will describe it within the model of Projective Simulation discussed in Chap. 5. For this we will assume that the manifold of references and involvements that underlies the agent's circumspection is formed by the individual sensory and motor experiences that the agent encounters during the practice and learning phase. These experiences are refined over time, until they become part of the background and are integrated

into the (action-guiding) structure of the PS agent's episodic memory. But even when they are no longer fresh in the agent's memory, they form the background of experience that guides and accompanies the skilful execution of hammering. In Sect. 6.5 we will describe how the development of a background of sensori-motor experiences could be mapped into the clip network of the PS agent.²⁰ This can be seen as a first step of modelling the being-in-the-world of a PS agent.

6.4.1 *Learning How-to vs. Learning What-for*

Learning *how-to* should be distinguished from learning *what-for*. Learning how to drive a nail into the board appears within a well-specified context and task—for example, as an apprentice of a carpenter business. Of course, one can use the hammer also for other purposes, for example, to temper iron or to tap a wall when looking for voids. For each of these purposes, the way how to use the hammer properly will vary, but it can be learnt in each case. Learning about the different uses of the hammer is different from learning *how to* use the hammer properly for a specific purpose, even though both types of learning are related in that they refer to ways how *Dasein* encounters and deals with equipment. In both cases, learning has to do with how the manifold of references and involvements that is constitutive of equipment changes over time, but in the case of learning *what-for*, that manifold may expand considerably.

The referential whole of equipment is thus not given as something finished and static. It is formed by the agent by learning about its different ends, which are often defined by social practices, and about its best means for each of these ends. But the referential whole also changes when the agent finds or even creates entirely new ends and contexts for them. The latter case is especially relevant in situations where *Dasein* plays around and tries out new things with entities it encounters, without a specific goal or task. Situations of this type are particularly important in the context of child development, but also in basic science, and so these situations are especially relevant in the context of modelling scientific agency that we discussed in Chap. 4.

The type of learning that plays a role in such situations—trying out new things and playing around with equipment—does not feature explicitly in *Being and Time*. It may not be altogether alien to Heidegger. Dreyfus (2007, 1146) highlights a

²⁰ This background of sensori-motor experiences can be related to the perceptual-motor field introduced by Merleau-Ponty (1945), in terms of which skilful hammering can be seen as a state of equilibrium:

Whether a system of motor or perceptual powers, our body is not an object for an 'I think', it is a grouping of lived-through meanings which moves towards its equilibrium. (Merleau-Ponty, 1945, 153)

certain developmental aspect in Heidegger's analysis by bringing to attention the following quote from his early works (for the notion of caring see Sect. 6.6.1 below):

Caring takes the form of a looking around and seeing, and as this circumspective caring it is at the same time [...] concerned about developing its circumspection, that is, about *securing and expanding its familiarity* with the objects of its dealings. (Heidegger, 2002, 115, italics by Dreyfus)

But the examples that Heidegger discusses typically refer to situations from rural life, where the meaning of equipment (its in-order-to, reference, and involvement) can be seen as static and as rooted in established social practices. Such examples concern the furniture in a room, lighting installations in public spaces, or equipment in the workshop of a carpenter. Similarly, in his abstract analysis, Heidegger (1927, 98 [69]) writes that dealings with equipment subordinate themselves to the manifold assignments of the “in-order-to”, which also suggests a static interpretation.

Another aspect with a static connotation is the closedness of the manifold of references, as highlighted by Heidegger (1985, 186f.): “This referential contexture is itself a *closed totality*”. One might argue, supporting Heidegger's view, that the manifold of references needs to be at least locally closed, defining a local background of what is already understood that is necessary for learning. Learning would then mean to connect different local manifolds, which are thereby broken up, updated, and expanded. But this first has to be accomplished, e.g., by showing in more detail how one can think consistently about (or even model) referential manifolds to begin with. Overall, it seems that learning in the form required for agency is not explicitly addressed in *Being and Time*.

In any case, from the perspective of agency, we must hold on to the fundamental role of learning. As stated in Chap. 2, learning is necessary for agency. But if we want to understand agency, in its basic form, as circumspect coping—the thesis we defended in Sect. 6.2.3.2—then this circumspect coping must accommodate the possibility of learning. It seems therefore natural to demand that in each case the concrete in-order-to, that is, the local context of references that constitutes equipment, must first be learned and established.

Equipment for Heidegger always comes with its “in-order-to”, which an agent can learn while she familiarises herself with it, possibly in a social setting. But sometimes things are used in novel ways, or entirely new things are discovered. What is the status of such entities—indeterminate equipment—that do not yet have an in-order-to, a determined reference? This question raises a challenge for the ontological fundamentality of the status of equipment. In any case, the status of newly encountered entities needs to be distinguished from the status of the mere occurrence (presentness-at-hand) of things. Otherwise, occurrent things that are present-at-hand would be ontologically prior to available equipment that is ready-to-hand. The context of references that constitutes equipment would then be something *added* to mere occurrence. This would undermine the entire methodological framework that starts from *Dasein* as being-in-the-world, rather than from the world as the totality of occurrent things. It is therefore unclear how learning of the type needed for agency finds its place within an orthodox Heideggerian perspective.

6.4.2 Indeterminate Equipment and Toys

On our approach, indeterminate equipment—equipment with an indeterminate status in terms of its in-order-to, involvement, and significance—comes *before* equipment as part of a finished and closed manifold of references. At the same time, it is to be distinguished from the occurrent thing. Occurrence in the sense of being present-at-hand remains ontologically secondary. Even if the relevance of each piece of equipment needs to be learned, *Dasein* does not start from occurrent entities. The child encounters entities in its environment neither as occurrent objects with fixed properties, nor as equipment with a finished structure of relevance. In his theory of perception and affordances, Gibson describes a similar situation as follows:

There is much evidence to show that the infant does not begin by first discriminating the qualities of objects and then learning the combinations of qualities that specify them. Phenomenal objects are *not* built up of qualities; it is the other way around. The affordance of an object is what the infant begins by noticing. The meaning is observed before the substance and surface, the color and form, are seen as such. An affordance is an invariant combination of variables, and one might guess that it is easier to perceive such an invariant unit than it is to perceive all the variables separately. It is never necessary to distinguish *all* the features of an object and, in fact, it would be impossible to do so. Perception is economical. (Gibson, 1986, 134)

To better understand the problem of learning from a phenomenological perspective, we need to consider situations in which learning plays a prominent role, as in child development, or, more broadly, in the *Dasein* of the child. On this topic, however, one cannot find much in *Being and Time*. The individual *Dasein* in *Being and Time* has almost exclusively the character of an *adult* human. The exemplary situations that Heidegger describes mostly portray the trained (“learned”) craftsman and his skilful and seamless, smooth dealing with equipment. The concept of skilful coping as discussed by Dreyfus (1991) evokes a similar image.

However, dealing with the world also includes the possibility of true discoveries and of creating something new. The child’s way of encountering and experimenting with environmental entities is in some respects closer to that situation than the apprentice’s attitude. In its playful dealing with (proto-)equipment, the child does not necessarily learn conventional uses. Instead, it tries out all sorts of things it can possibly do with it. It seems appropriate to use the word *toys* for equipment that is undetermined with respect to its ends and involvements.

For an extended ontological framework that can account for the type of learning that is needed for agency, we thus have the ontological hierarchy of *toy–equipment–thing*. In that framework, we need to assume a kind of indeterminate readiness-to-hand for toys as (proto-)equipment. This will connect directly to the options for action in the PS model, to be discussed in the following Sect. 6.5. Whether such indeterminate readiness-to-hand refers to natural entities or artefacts is not important. As an agency-theoretical structure it applies to agency across

the board, including the agency of simpler animals, for which affordances and (proto-)equipment seem to largely coincide.²¹

6.5 A Phenomenological Interpretation of Projective Simulation

Having laid out the background of a phenomenological approach to agency and its specific challenges, we are now in a position to consider the agency model of Projective Simulation (PS) discussed in Chap. 5 from a phenomenological perspective. Our discussion in this section has three main targets. First, we demonstrate how the PS model underwrites our main thesis from Sect. 6.2.3.2: *agency, in its basic form, is circumspect coping*. This includes a discussion of the basic agent-environment framework of a PS agent under a phenomenological reading. Our second target is to describe the development of smooth coping from the point of view of PS, using the example of an agent who learns how to hammer. This includes a (toy) model of the referential manifold that characterises circumspect hammering, followed by a more general discussion of circumspection as it plays out in the PS model. Third, we show how thrownness and projection, two of the existential temporal dimensions of *Dasein* discussed in Sect. 6.3, fit into the PS framework. In the following Sect. 6.6, we will then present some thoughts on how simple forms of self-reference and care, two key aspects that are considered specific to human agency, can already be located in the PS model, and we will show how the PS framework can provide steps towards an account that unifies an embedded and a detached view of agency.

6.5.1 Preliminaries and Basic Framework of PS and Reinforcement Learning

We refer to Chap. 5 and specifically to Sect. 5.2 for the technical background of PS. It will be useful to start our phenomenological interpretation by considering parallels between PS and the general framework of reinforcement learning, as discussed in Sect. 5.3. One important point here is that already in reinforcement learning

²¹ The relation between affordances and equipment certainly deserves further investigation. In the context of our analysis from Sect. 6.2.3, an iterative, two-step argument could be useful. To a first approximation, both affordances and equipment are seen as static structures: While the referential structure of affordances is determined by the species, the referential structure of equipment is socially determined. To a second approximation, however, neither of these structures is completely determined. The individual animal agent needs to use and explore affordances, especially while the animal develops and learns. The individual human agent, in particular the child, explores the referential structure of (proto-)equipment (toys) and may even establish new connections and uses. There is probably more room to manoeuvre in the human case. But still, there is a strong similarity.

considered generally, the environment plays a constitutive role for the agent. An agent is always defined relative to an environment, it is *situated* in an environment. Like a generic reinforcement learning system, a PS agent encounters its environment in the form of percepts $s \in S$, and it can act on the environment through actions $a \in A$. In PS, environmental objects are typically not explicitly specified (there is no “knowledge base” or similar). Rather, objects exist for the agent only insofar as they can be encountered through percepts, actions, and rewards.

In reinforcement learning, $s \in S$ is typically referred to as “the state of the environment”. If one interprets s in this way, then the transition function $T(s, a, s')$ introduced on page 275 (Eq. (5.13)) describes how the environment is changed by the agent’s action a . As discussed in Sect. 5.3, in the popular mathematical model of a task environment known as a Markov Decision Process, one assumes the existence of certain mathematical functions $T(s, a, s')$ (state transition function) and $R(s, a, s')$ (reward function), which together define the environment in which the agent acts. With a slightly different focus, s can also be read as the state the agent *finds itself in*: If the agent chooses an action a in state s , then it will (with a certain probability) find itself in a new state s' .²²

The boundary between agent and environment is not necessarily fixed and can be drawn in different ways depending on the application and context. This openness and flexibility in drawing the agent-environment boundary is a noteworthy feature of the reinforcement learning framework. It is best to think of the state s as assigned to both the agent *and* the environment, or to the agent-in-the-environment. This fits well with our phenomenological approach in which we adopt being-in-the-world as the basic characterisation of agency. As mentioned above, being-in-the-world is seen as primary, while a division into a subject and its world of objects is something that needs to be derived and established separately—if wanted and warranted.

It is not necessarily assumed that the agent represents, let alone knows the transition function T of the environment. Reinforcement learning without this assumption is called *model-free reinforcement learning*, in contrast to *model-based*

²² Such an interpretation is especially fitting for the *grid world* task environment discussed towards the end of Sect. 5.2.4, where an agent moves in a maze considered as a deterministic environment. In *grid world*, s corresponds to a label of the room in which the agent finds itself, and a corresponds to exiting the room through one of several doors, after which the agent finds itself in a new room. The task environment described in this way allows and even suggests a view *from the outside*, from where we can watch, so to speak, the agent wander around in the maze. In that case, one can infer from $T(s, a, s')$ the topology of the maze, i.e., the objective structure of the environment itself. Such a view from the outside is often implicitly presupposed. This is strikingly obvious when one considers the terminology used to describe *Partially Observable* Markov Decision Processes, where the agent has only partial information about the state s it encounters. In such task environments, a distinction is made between the states s themselves and those aspects of the states that are observable for the agent (observations o). In our PS terminology, *these* would be the percepts, and a view from the outside is not needed nor presupposed. (In fact, it is not needed for the basic layout of reinforcement learning either, as we emphasised earlier.) At any rate, an outside view reaches its limits when we consider quantum mechanical laboratories as task environments, examples of which were discussed in Sect. 5.4.

reinforcement learning where the agent may know or estimate and then make use of this function (see Sect. 5.8.1). Without any knowledge of $T(s, a, s')$ that could guide the agent, it can only try out and see what happens. The agent is *thrown* into its environment, as it were, without further advice. As we pointed out, Heidegger describes thrownness as an *existentiale* of human agency (*Dasein*), but the notion has broader application. It describes the basic situation of an agent finding itself in a world that it has to cope with, without further advice or specific instructions on how to go about. A minimal feature of thrownness is thus built into the very framework of reinforcement learning, and similarly into Projective Simulation.

In model-based reinforcement learning, the agent tries to learn the transition function as a model of the environment. This may be useful, in particular, for tasks like planning, and it may help the agent to find an optimal policy, i.e., a behaviour that maximises the long-term expected return, if there is such a thing. From a phenomenological perspective, it may be tempting to interpret the transition function $T(s, a, s')$ as the fundamental referential manifold associated with the environment, which would then enable circumspect coping. This would imply that the referential context is somehow fixed and objectively given by the environment, rather than something to be created and developed by the agent as it learns to circumspectly cope with equipment. For the problem of skill learning, i.e., learning *how to* skilfully cope with given equipment, or how to exploit the affordances provided by a biological environment, such an interpretation may be possible. For other problems, however, this static view is too restricted and will need to be refined. For example, a static view is too narrow when the set of states and actions is not fixed, i.e., when that set can vary or may not even be determined at a given time. We mentioned a relevant concrete example in Sect. 5.4.3, where we indicated how a PS agent can learn how to correctly predict the orientation of a spin-1/2 particle in an external magnetic field by a refinement of the repertoire of possible measurements on the particle, and how to adapt its actions through action composition. More generally, a static view is inappropriate in the context of learning of the *what-for* type, where the different uses of (proto-)equipment are undetermined and have yet to be established.

6.5.2 *Specific Aspects of the PS Framework*

The PS model exhibits the same minimal notion of thrownness as reinforcement learning. At the same time, it goes beyond reinforcement learning in two different ways, both of which are relevant for its phenomenological interpretation. One of these ways concerns the feature of clip composition and creation. This feature provides an important refinement in the description of learning of the *how-to* type, and a possibility for tackling learning of the *what-for* type. We will discuss these cases in Sect. 6.5.3. The second way in which PS goes beyond reinforcement learning concerns the feature of projection. This feature builds upon the PS agent's Episodic and Compositional Memory (ECM) as a simulation platform where

the agent simulates its past experiences in dealing with the environment. The ECM is an embodied platform, built from a homogeneous set of components, namely, remembered percepts and actions that describe basic encounterings, without importing additional external computational machinery. There is no ghost in the machine and no external advice. Projection in PS is thus based entirely on the agent's embodied structure that is part of its thrownness. Projection is a key property of the PS model, in particular under the phenomenological reading, where projection constitutes another important *existentiale* of (human-like) agency. We will come back to this in Sect. 6.5.5.

Of course, as in any formal model, simplifications are made. In its basic layout, the agency model of PS conforms to the reinforcement learning framing of agent-environment interactions in terms of discrete rounds that start with a (given) percept and result, via a deliberation whose internal timing is normally not resolved, in some action. In Chap. 5, we pointed out that PS goes beyond this basic model in offering an internal dynamics of deliberation via the hopping of an excitation, triggered by a percept clip, within the ECM. Furthermore, in discussing quantum PS in Sect. 5.6, we already discussed further ways of loosening the basic structure, e.g., by allowing for multiple excitations.

With a view to integrating insights from phenomenology, it indeed seems important to remove the idealisation of a discretely timed percept-deliberation-action structure. Agents receive input from their environment as a constant stream of impressions, with differing time scales for different sensory modalities, and actions are often executed under continuous feedback, e.g., from proprioception. Such considerations will certainly be important for a fully versatile PS agent, where one may envisage multiple excitations within an ECM in a continuous stream of smooth sensor-motor integration. It should not be forgotten, however, that modelling is necessarily tied to some forms of idealisation, and that this is fundamentally a good thing: An idealised model can be executed and studied in detail, while an attempt to put everything into a comprehensive model may result in failures that cannot even be diagnosed. So while we fully acknowledge the limitations of the round-based percept-action scheme as a distortion of the temporal structure of absorbed coping, we will stick with this idealised modelling for the time being in order to keep things manageable.

At the same time, we extend the basic reading of s and a as percepts and actions towards a broader interpretation as individual (sensori-motor) *encounterings* of the agent in dealing with the environment. Clips c and c' then encode previous sensori-motor encounters of the agent with the environment. The clip network of a PS agent differs from the transition function $T(s, a, s')$ of the environment. For example, transitions between different clips need not involve an action a ; furthermore clips can encode “fictitious” encounterings that never happened. The clip network is the central structure of a PS agent. It is different for each agent, because it is shaped by the agent's individual learning history, and it gives the agent a form of identity. This identity also determines its specific way of dealing with equipment and what it takes to be significant. It is important to note that the clip network does not describe the outside world as such. Rather, it describes an agent-relative action context in which,

for example, certain action sequences were rewarded in the past and proved to be *good* for the agent in some sense. On our phenomenological reading, we will thus interpret connections from a given clip to other clips in the ECM as *references*, and the clip network as the basic referential structure of a PS agent's dealing with its environment.

It is clear that the interpretation of the ECM as a network of references comes with the danger of levelling and oversimplification. The references of equipment can be arbitrarily complex, more complex than what can be represented via simple graphs. Equipment refers in manifold ways. But the PS model's ECM is a good starting point to explore these issues, and there are various possibilities of refining and extending the PS model, as mentioned—for example by considering multiple excitations—that can be included when needed.

6.5.3 An Example of Circumspect Coping

As an illustration, we reconsider the example of a human action—driving a nail into a board with the help of a hammer—and the description of a skill—hammering—that we gave at the beginning of Sect. 6.4. To model this example with Projective Simulation, we take the clips to describe snippets of sensori-motor experiences that the agent encounters in the process of learning how to hammer. Examples of such snippets could be: “nail not gripped tightly”, “hammer too heavy”, “nail hit sideways”, “nail flies away”, “pain in finger”, etc., as will be discussed in the context of Figs. 6.2 and 6.3. In contrast to the applications of PS discussed in Chap. 5, where

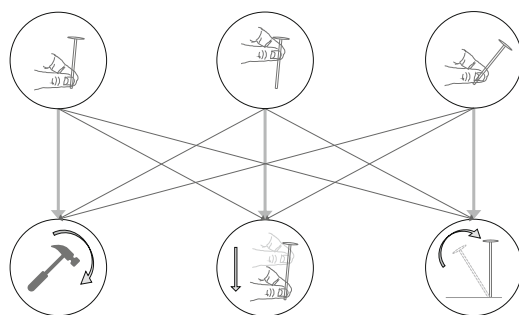


Fig. 6.2 Learning *how-to*. The figure shows part of the referential manifold associated with hammering. Percept clips (upper layer) correspond to a nail held in different ways. These are connected to different action clips (lower layer). For example, the right clip in the upper layer describes a situation where the nail is held in a misaligned way. The possible actions displayed are: hit the nail on the head with the hammer (lower layer left), adjust the position where one holds the nail (lower layer center), adjust the alignment of the nail (lower layer right). Learning how to skilfully hammer leads to larger weights (probabilities) associated with certain edges. This is indicated via thicker lines of the edges that refer to the appropriate action in each case. For the upper right clip, for example, the appropriate action is to align the nail *before* hitting it with the hammer

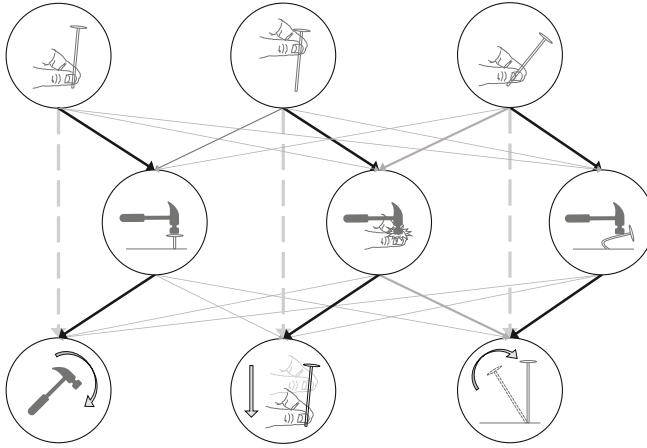


Fig. 6.3 Learning *how-to* hammer, refined. Clips in the middle layer correspond to different encounterings during the learning process. These encounterings shape the way the agent learns how to cope circumspectly with the hammer. The resulting manifold is constitutive for the hammer *as equipment*. For example, the nail held in a misaligned way (upper layer right), as the initial perceptual situation, points to a remembered encountering of a crooked nail (middle layer right) which, in turn, points to the action “align the nail” (lower layer right) that the agent then performs. Depending on how skilfully the agent performs the alignment, the next perceptual situation could correspond to the clip in the center of the upper layer where the nail is held upright, near its head. This clip points, e.g., to a remembered encountering “finger painfully hit” (middle layer center) which, in turn, points to the action “hold the nail near the tip” which the agent then performs, etc. Of course, each of the clips in these examples refers also to other clips, describing different associated encounterings, which may be less fitting and lead to unsuccessful performance of the hammering. But as the agent learns to hammer successfully, the weights (indicated by the thickness of the edges) in the referential manifold will adapt accordingly. The Figure indicates the finished result of such an adaptation. Dashed arrows, pointing from the upper to the lower layer of clips, indicate direct references that may build up and dominate the other, fading connections over longer time scales, depending on the implementation of the learning rule in PS

clips throughout carry high-level semantics given by the task environment, the semantics of the clips in skill-learning such as in our hammering example is lower-level. Deliberation here corresponds to a sequence of sensori-motor impressions. However, the action-guiding and action-forming function of the clip network is the same in both cases. Similar as in the higher-level examples, one can here specify a series of sensori-motor impressions that can be re-invoked during deliberation and thus prepare the hammer blow. The indeterminism in the transitions between the clips is essential also in this case; it opens the room for the agent’s projections by offering different deliberation paths.

Figure 6.2 exemplifies the referential structure associated with hammering. The agent is a carpentry apprentice, and the initial situation is this: A nail held on the board with the objective to drive the nail successfully into the board—that is, to hit it hard and accurately and not to miss it. This initial situation refers to different possible actions: “hit with the hammer”, “hold the nail at another position”, “adjust

the alignment of the nail". Depending on the more precise specification of the initial situation, one of these actions is the appropriate one, and after a successful learning process, the agent will perform the respective action with high probability. In the corresponding clip network shown in Fig. 6.2, the percept clips of a nail held in different ways, as they occur exemplarily in the learning history of the agent, refer to different action clips, and the strongest connections exist to the appropriate actions. Here, each percept clip is directly connected to different action clips. Even such a simple, shallow two-layered clip network can illustrate the referential manifold behind smooth coping. Note that the experienced carpenter is indeed characterised by the fact that she does not have to think much when dealing skilfully with the hammer (her equipment); the more she is engaged in her activity and absorbed by it, the less she thinks.²³ In this simple two-layered network, the effect of learning *how-to* can be represented as a change in the weights of the transitions between percept and action clips. However, deeper clip networks can account for more subtle aspects of sensori-motor learning. Figure 6.3 shows a three-layered network in which the phase of learning of how to skilfully hammer can be mapped out in more detail. That network also accounts for episodic experiences that represent different, more or less successful ways of holding the nail, along with the painful hit on the finger or the crooked nail.

The refined representation of learning how to hammer in Fig. 6.3 illustrates the concept of *circumspection* that we already discussed above. In a situation where the agent is about to hit the nail while holding it at an angle, the corresponding clip (upper layer right) in the referential ECM points to several situations that the agent has encountered before, and which could come to its mind, as it were, in the process of dealing with this situation. These include sensori-motor experiences where the nail was crooked (middle layer right), hitting the finger caused pain (middle layer centre), or the nail was successfully driven in. These experiences, which are all part of the referential whole associated with the hammer(ing), play together and provide the specific local *sight* and the more encompassing *understanding* that guide the action and that enable circumspect coping. Note that the individual sensori-motor experiences that are collected during the practice and learning phase are an integral part of the learning process and part of the constitution of the action. Over time, the transient experiences made during the learning process are integrated into the background of references and involvements. But over time they may fade and be masked by experiences corresponding to direct transitions as shown in Fig. 6.2. This fading can be modelled in PS by a process where the weights of the respective transitions in the ECM weaken and the direct connections between percept and action clips are enhanced, as described in more detail in Briegel and De las Cuevas (2012, pp. 8f. and Figure 10). Nevertheless, the indirect connections still provide a background of references which contribute to the circumspect handling of the hammer.

²³ "[...] the less we just stare at the hammer-Thing, and the more we seize hold of it and use it, the more primordial does our relationship to it become, and the more unveiledly is it encountered as that which it is—as equipment." (Heidegger, 1927, 98 [69])

As discussed in Sect. 6.4, learning how to skilfully hammer a nail into a board is an example of learning *how-to*. More complex forms of activities may also involve learning *what-for*. Examples of such activities that involve a hammer could be building a hut (in order to get shelter) or breaking a window (in order to quickly escape from a fire). In these examples, circumspection is provided by a referential structure with clips on a different level of resolution than in the case of skilful coping. For example, the hammer in the agent's hand refers to the nail, which refers to the wooden scaffolding, which refers to the roof-over-the-head, which refers to being protected from the weather, for the sake of which the agent ultimately wants to build the hut. Figure 6.1 shows the corresponding referential manifold, together with other connections that together span the space of significance associated with the hammer. Clip networks can represent such referential structures and their development in both types of learning scenarios; clip creation and composition can be employed to represent novel uses of the hammer that the agent may discover (or create) in different contexts, representing different *what-fors*. This allows us to model agents and their being-in-the-world as circumspect copers as well as the trajectory of learning that leads them to develop their respective skills.

6.5.4 Developing Circumspection

Informally, a clip network describes what can possibly come to the agent's mind in an encounter with its environment, i.e., in a given perceptual situation and with regard to an action to be taken. "What can possibly come to mind with regard to an action to be taken" is not meant in the sense of a description of mental states, but it describes circumspection as a basal form of awareness (or seeing). Figure 6.3 illustrates this for the example of hammering. In the ECM, the sight of circumspection covers, as it were, what lies within the accessible range (or the "light cone") of a perceptual clip under the dynamics of deliberation, and can thus be excited by the Projective Simulation that provides understanding and prepares and guides the action.

Circumspection is thus enabled by the action-guiding simulation making use of the referential structure represented as the weighted graph of the clip network. This referential structure also establishes an individual *perspective* that the agent brings *with it* when encountering equipment. That perspective corresponds to the fact that the clip network—both in its topology and in the weights of the transitions—develops with the agent. It does not represent the world *per se*, but the world *for the agent*. This grounds a first-person view that changes and adapts over the course of the agent's learning history.

In the previous Sect. 6.4 we discussed different ways of learning and thus of developing circumspection. This included learning *how-to* and learning *what-for*. In the ECM the first way is primarily described by a change of the weights of the existing referential manifold (the weighted graph of the ECM), while the second way is described by a change of its topology, e.g., by the addition or composition of

clips and edges between them. Fig. 6.2 and 6.3 illustrate these two different ways of developing circumspection for the example of hammering.

Both types of learning play a role in applications of PS to scenarios in basic science, some of which were discussed in Sect. 5.4. For example, the application of PS to learning the behaviour of a spin-1/2 system under the influence of a magnetic field, as discussed in Sect. 5.4.3, exemplifies learning *how-to*. It involves the adaptation of weights and the composition of clips and thereby provides the structure behind smooth coping. The application of PS to the design of novel optical experiments as discussed in Sect. 5.4.4, on the other hand, mostly corresponds to learning *what-for* and involves the creation of new clips.

Through the study of these applications, we have seen how the different ways of learning feature in the PS model. The phenomenological interpretation of the PS model, where clips generally encode encounterings with equipment, naturally leads to the question of how to interpret clip composition and clip creation. It also raises the question of the ontological status of new equipmental clips, which can be added to the network in different ways. Different ways of extending a given graph lead to different meanings for a newly added clip, defining its *in-order-to* and thus its specific role, if any, as equipment. Newly created clips may often best be associated with equipment without a prior *in-order-to*, i.e., with *toys* (see Sect. 6.4.2).

The phenomenological interpretation of PS thus meets the first target listed at the beginning of this Sect. 6.5, viz., the underpinning of our thesis that agency, in its basic form, is circumspect coping with equipment. It furthermore shows that learning—a basic element of agency—requires the modification of the ontological status of equipment, but it also suggests a solution. As a result, the integration of learning into the existential structures that describe *Dasein*, including its development, affirms the ontological hierarchy of *toy–equipment–thing* that we posited at the end of Sect. 6.4.2.

Next we will discuss a phenomenological interpretation of the PS framework with respect to projection and thrownness. These *existentials* highlight intrinsic temporal dimensions of *Dasein* and its possibilities. Together with absorption, thrownness and projection can be seen as the key elements of being-in-the-world. Dreyfus and Taylor (2015) refer to these elements in their succinct summary of the (human) agent as an *open discloser of a world*:

Heidegger describes this threefold structure of being-in-the-world as thrown, absorbed (falling), and projection, and shows how this structure enables *Dasein* to open up a world. As the being that discloses a clearing or world, *Dasein* has no essential structure other than the basic structure of being-in-the-world. *Dasein*'s essential structure, then, turns out to be that it is an open discloser of a world. (Dreyfus & Taylor, 2015, 161)

6.5.5 Projection

In the PS model as described and used in Chap. 5, the ECM functions as a simulation platform on which experiences that the agent has had in similar encounterings are

replayed and varied. This replaying and varying is not to be understood as conscious thinking, but as a basal simulation process, as already discussed in Sect. 5.1. This simulation process takes place in-between different encounterings and allows the agent to project itself into future situations that are plausible in view of its individual learning history. The projection, in turn, does not have the character of calculating and planning the best action to reach a given goal, as would be conceivable with the help of an internal representation of the outside world, but rather that of drafting and sampling action possibilities in the given referential context represented by the clip network. As we pointed out in Sect. 5.1, we use simulation and projection in a basal sense, which then informs a basic notion of deliberation:

- *Simulation* means to go through something similar. That is, going through episodic sequences that start with the same perceptual impression but may vary in terms of the order and content of the clips that are excited subsequently.
- *Projection* means to throw, or to run ahead. That is, running forward along episodic memory sequences to arrive at the possible actions that could be taken.
- *Deliberation* then denotes the whole random process of forward-running simulation that starts with a percept and ultimately ends with an action.

In the PS model, the agent thus does not simulate the outside world. By re-invoking episodic memories (clip sequences) associated with a given encountering, the agent goes through something that is similar to past encounterings. It is thus more appropriate to speak of the PS agent simulating *itself*, that is, its own possibilities in each situation. Like *Dasein*, the PS agent projects itself into open possibilities. Furthermore, these possibilities are always individual possibilities, i.e., strictly the PS agent's own.

As mentioned in Sect. 6.3.3, projection into possibilities is the agent's way of *understanding* the situations into which it is thrown. Understanding is meant in the sense of seeing possibilities via circumspection, not as thematic consideration. Taken all together, the PS model thus naturally fits with the phenomenological idea of projection and with its constitutive role for understanding and circumspection.

6.5.6 Thrownness

Thrownness describes the basic situation of an agent finding itself in a world that it has to cope with, without further advice or specific instructions for how to go about. For artificial agents, what that basic situation looks like depends to a large extent on the designer. This in turn strongly influences how well the agent is equipped for the situation it is thrown into, and to what extent it can be successful in coping.

As described earlier, one of Dreyfus's (1992) main criticisms of symbolic AI was in fact that this approach to AI would necessarily fail when viewed as a method for designing an artificial agent. This critique was reflected in actual developments of robotics starting in the 1980ies. For example, Brooks (1991) argued that the field should move away from methods based on knowledge representation and planning

towards more embedded and so-called situated approaches (Clancey, 1997). As pointed out in Sect. 6.5.1, PS and the framework of reinforcement learning fit well with the phenomenological idea of thrownness. The very design of an artificial agent includes the types of sensors and actuators it is endowed with, which defines the ground-level situation for the agent under which it has to operate. Formally, this situation is marked by the state space S and action space A . In that sense, as pointed out above, a reinforcement learning agent is *thrown* into an environment specified by S and A . In any state $s \in S$, it can take various actions $a \in A$, and it has no further guidance such as externally provided advice. The only feedback comes in the form of the reward, and that reward ultimately has an effect on the agent's physical condition.²⁴ The PS framework adopts this minimal notion of thrownness, but it goes further and integrates it with other structural moments of being-in-the-world, including circumspection and projection as discussed in Sects. 6.5.4 and 6.5.5. This integration is achieved in at least two ways.

First, the sets of percepts and actions that describe the agent's possible ways of encountering its environment also directly enter its episodic memory (ECM), since the clips that constitute the ECM are defined over S and A (see Sect. 5.2, esp. Sect. 5.2.1). This means that the thrownness associated with the agent's ground-level situation described by S and A —its basic way of *being-in*—*carries over* into its central structure, the ECM, which grounds the agent's capacity of circumspect coping and projection. There is nothing more. The PS agent is thrown through and through.

Second, in the PS framework, projection and thrownness are closely tied together in yet another sense. Any encountering that starts with some perceptual input triggers a random process in the agent's ECM, which amounts to a projection. (That projection could even be seen as part of the encountering.) Furthermore, any change of the ECM due to new factual experiences will feed back and change the way the PS agent projects. This corresponds to the way in which learning and deliberation are connected and play together in PS (see Sect. 5.2). It therefore seems that the only way the PS agent, as a thrown entity, can be an agent is to project.

This close connection between the agent's circumspect absorbed coping with a situation into which it is thrown, having to respond to the next perceptual input, and its continuous projection into possibilities resonates well with the discussion of thrownness and existential possibilities in Sect. 6.3.

6.6 *Dasein* and Scientific Agency

In this Chapter we have developed a phenomenological perspective on agency. Our focus was on a low-level notion of agency that can apply not only to *Dasein*—human

²⁴ The feedback on whether an action taken by the agent was good for it may come with some delay and is usually unspecific, i.e., without further information as to which part of a longer chain of actions was decisive or most relevant.

agency—but also to animals and to artificial agents. As shown in Sect. 6.5, our approach is supported by a phenomenological interpretation of the PS model. Due to that model’s explicit formal, low-level nature, our phenomenological interpretation offers a way to read the *existentials* of *Dasein* in a minimalistic fashion.

This perspective can be developed in at least two ways, both connected with the difference between general agency and human agency (recall our distinction of different types of attribution in Sect. 2.3.2). On the one hand, it can help to better position ourselves, as human agents, in the ongoing technological development of AI that may soon lead to artificial agency. On the other hand, it opens up a space for a more general discussion of the specifically human aspects of agency as exhibited in the analysis of *Dasein*: Given the low-level anchoring of the phenomenological framework that we developed by transferring the notion of being-in-the-world downwards, as it were, can we find ways to reach back to *human* agency?

While we cannot discuss this question in its full complexity, our approach provides an interface that can be exploited and further developed in different directions. In the following we sketch two possible directions. First we carry the phenomenological analysis of agency a bit further, towards self-reference, care, and temporality. Then we focus on scientific agency, which includes the development of concepts and models and thereby connects the phenomenological analysis to our discussion in Chap. 4.

6.6.1 Self-reference and Care

As pointed out in Sect. 6.3, one of Heidegger’s key notions in the analysis of *Dasein* is the self-referential structure of *care*, which describes a specifically human trait that need not be present in all agents. Care stands out as one of the existential dimensions of *Dasein*, and it is an interesting question how one could possibly model this dimension in PS agents. Some aspects of care can already be found in the *existentials* of absorption, thrownness, and projection, all of which refer to forward-pressing aspects of agency, including PS agency. Specifically, absorbed coping is enabled by a continuous projection into open possibilities.

Phenomenologically, care adds a strongly self-reflective element to these forward-looking aspects. This element is related to the simpler feature of *wanting* that we discussed in Sect. 2.6.1: Wanting means that there has to be something *in it* for the agent that gives meaning to its actions. Drawing an (admittedly highly simplified) parallel, one could say that *care* refers to some feature of the agent that gives meaning to its own being. Since human agents are fundamentally characterised by their possibilities (Sect. 6.3), the ultimate meaning of the agent’s agency—and thus what it is ultimately concerned with—is to maintain open possibilities. And the way how an agent maintains its possibilities is by continuously projecting into new possibilities.

In his analysis of the structure of care, Dreyfus (1991, 238) writes that “care unifies the various structural aspects of *Dasein*’s way of being”. In support, he quotes the following passage from Heidegger’s *Being and Time*:

Dasein exists as an entity for which, in its Being, that Being is itself an *issue*. Essentially ahead of itself, it has projected itself upon its potentiality-for-Being *before* going on to any mere consideration of itself. In its projection it reveals itself as something which has been thrown. It has been thrownly abandoned to the ‘world’, and falls into it concernfully. As care—that is, as existing in the unity of the projection which has been fallingly thrown—this entity has been disclosed as a “there”. (Heidegger, 1927, 458 [406])

In *Being and Time*, the structure of care is disclosed by a detailed analysis of *anxiety* as a distinctive mode of *affectedness* (*Befindlichkeit*). In Dreyfus’s words: “To reveal *Dasein* simple and whole Heidegger chooses anxiety. Just as the breakdown of a piece of equipment reveals the nature both of equipmentality and of the referential whole, so anxiety serves as a breakdown that reveals the nature of *Dasein* and its world” (Dreyfus, 1991, 177).

In Heidegger’s analysis, anxiety refers to a crucial and specific feature of human existence which we are not concerned with in our analysis of agency, and which has no obvious place in the PS model. Thus in our approach, we cannot represent anxiety as that which reveals care—but this does not mean that we have no means to represent the structure of care. Indeed, in the formal characterisation of care as given in the above quote from *Being and Time*, there is no explicit reference to anxiety, nor to authenticity or inauthenticity, even though these also play an important role in Heidegger’s analysis (see Sect. 6.3.3). The minimal formal definition of care that can be extracted from the above quote summarises *Dasein* as care in the sense of an entity that exists “in the unity of the projection which has been fallingly thrown” (Heidegger, 1927, 458 [406]).

This also resonates with Dreyfus and Taylor’s characterisation of *Dasein* as a “threefold structure of being-in-the-world as thrown, absorbed (falling), and projection” (Dreyfus & Taylor, 2015, 161), which we already showed to be exhibited by the PS framework.²⁵ The self-referential aspect of care may not be directly obvious from its minimal formal definition, but it is explicit in the characterisation of *Dasein* as “an entity for which, in its Being, that Being is itself an *issue*”, also quoted above. In a minimal sense, there is a self-referential element already in *the unity of projection*, insofar as the target of projection is not other entities in the world, but *Dasein* itself, as an absorbed and thrown entity.

Related to this characterisation, projection enables *Dasein* to be continuously *ahead of itself*:

Dasein is always ‘beyond itself’ [“über sich hinaus”], not as a way of behaving towards other entities which it is *not*, but as Being towards the potentiality-for-Being which it is itself. This structure of Being, which belongs to the essential ‘is an issue’, we shall denote as *Dasein*’s “*Being-ahead-of-itself*”. (Heidegger, 1927, 236 [192])

²⁵ See the longer quote from Dreyfus and Taylor (2015) at the end of Sect. 6.5.4.

The PS model exhibits this aspect of being-ahead-of-itself in a basic way. As we explained in Sect. 6.5.5, by projecting, the PS agent does not simulate the outside world, but its own agency, understood as its open possibilities in every given situation. For this it is not necessary that the agent's memory—the ECM—represent the agent *as* an agent in a thematic way. The ECM represents a condensed form of the agent's history and grounds its identity. One could also say that the ECM represents a *model of the agent* at any given time. The PS agent is then always ahead of itself because it has the ECM as a model of itself, but it exhibits its agency only as long as it projects itself on the basis of its own model.

Clearly this basic aspect of self-reference is not sufficient to account for the full existential meaning of care, and it is not clear how more human-like aspects of care, including social aspects, could be modelled. Such higher-level features may indeed require that the agent's ECM has the capacity to represent the agent *as* an agent *to* itself, in a thematic way. This may require linguistic or at least advanced conceptual capacities. These go beyond our current treatment, but some of them will be addressed in the next section.

6.6.2 *Developing Concepts and Models*

We have seen that, on the phenomenological interpretation, agents have a kind of understanding of their environment that is exhibited in skilful coping. This kind of understanding is provided by circumspection and projection. Traditional theories of knowledge typically discuss a higher-level type of understanding based on abstract concepts, models, and even theories. There are different views on what precisely concepts and models are, how they are related, and how to draw the line between the different notions of understanding. In Chap. 4 we characterised models as tools for understanding and getting around in the world. That characterisation fits well for scientific agents, but also for agents more generally. We observed that models as tools mediate between the agent and the world, pointing both ways. This pointing-both-ways is what models share with equipment and affordances. The structural analogy goes even further. In a similar way as affordances are economical—in the way their relevance is encoded for the agent—models as tools are economical as well. Economy is considered a hallmark of good scientific models, but it is also exhibited by models that are learnt as basic action-guiding structures in PS.

Models as tools have a basic action-guiding role, and they have a foothold in a phenomenological interpretation of agency. Like equipment, they are part of the referential manifold that constitutes circumspection and the kind of understanding that comes with it. Furthermore, at least in the context of scientific agency, it seems clear that models can also provide further insights into the environment and show how different phenomena are connected. Scientific models such as the Bohr model have a strong *unifying* character: first, they bring together different phenomena that previously may have seemed unrelated, and second, they enable the agent to conceive of and make predictions for new experiments. This capacity for unifying

different phenomena and providing insight into how they are connected is not just characteristic of scientific models; it is also attributed to whole theories or to the abstract concepts that inform them, such as mass, charge, and energy.

Within our phenomenological approach, with its systematic focus on developing agency from being-in-the-world, it may seem unlikely to find a place for models or abstract concepts. In cognitive science and in the philosophy of mind, the latter are typically seen as indicators of an agent's capacity of meaningful deliberative thought. However, as we have emphasised in Sect. 6.2, the view of actions as the results of a conscious subject's explicit mental deliberation plays no role in our analysis. In this we agree with Dreyfus (1991) and Heidegger (1927).

The question where one can find abstract conceptual structures in an embodied-embedded approach to agency has been the topic of a controversial debate between Dreyfus and McDowell (Schear, 2013). Within the PS framework, one can give examples for the appearance of structures that play a role similar to concepts and models. These examples may at least provide an idea how PS could be further developed towards a framework suitable for agents with higher and human-like cognitive capacities. They may also provide a fresh perspective on how the different positions presented in Schear (2013) might be reconciled.

We start with the discussion of models. As we already said, we consider models as tools whose primary role is to guide the agent in its actions. In the application of PS in quantum experiments described in Sect. 5.4.4, the process of learning how to design experimental set-ups was facilitated through *gadget formation*. Gadgets correspond to blueprints of little devices (in our example, interferometers) which were encoded via composite clips in the ECM. These clips were identified in a *post hoc* analysis of the clip network of PS agents that successfully generated several interesting experiments. We can leave it open for discussion whether gadget formation should be regarded as the development of a model, and whether gadgets are models. But in any case, gadgets correspond to characteristic structures of the clip network.²⁶ These structures develop under learning and help the agent to successfully deal with its environment.

In this example, but also more generally in the phenomenological interpretation of agency as being-in-the-world, models can thus be seen as *structures based on interactions that structure future interactions*. In future directions of modelling scientific agency, one can consider situations in which the agent makes use of these memory structures in a more flexible way. For example, a study by Trenkwalder et al. (2023) indicates how an agent could enhance, phenomenologically speaking, its radius of circumspection and its understanding of the environment beyond the specific context and task in which it first develops gadgets. It is interesting to explore whether gadgets should then qualify as structures that enable higher cognitive capacities, such as the use of concepts.

²⁶ In the application described in Sect. 5.4.4, gadgets are represented as composite clips, but they could also be conceived of as sub-networks of clips with nearly deterministic connections.

The investigation by Eva et al. (2023), already briefly discussed in Sect. 5.5.2, is pertinent here. The capacity of a PS agent to infer the existence of *unobserved variables* in its environment helps the agent to understand the behaviour of environmental objects encountered in experiments. The study shows how a PS agent can develop and utilise abstract conceptual structures, and what that means operationally. Eva et al. (2023) investigate how clips in a three-layered ECM network can acquire the role of abstract concepts, which can help the agent to employ its previous experiences in dealing with environmental objects in a new context. In the example discussed in that paper, the behaviour of different objects that the agent encounters in its experiments is determined by the value of certain variables such as mass, charge, and size. Knowing these variables and their values for every object would thus allow the agent to predict the outcomes of various experiments—for example, how much the object will deflect a compass needle when it is moved past it at a certain distance and with a certain speed. Eva et al. (2023) show how abstract concepts are formed and represented by the clip network and its connectivity, using a reinforcement-learning scenario where the PS agent is rewarded if it makes a correct prediction for different experiments. Identification of the structures within the ECM that represent those concepts requires a detailed *post hoc* analysis of the ECM after the agent has learnt successfully.

6.6.3 Agency: Both Embedded and Detached

It would be interesting to connect the considerations about abstract concepts or conceptual structures that we just presented with an *embedded* view of agency²⁷ and with the phenomenological interpretation that we developed so far. In closing this Chapter, we offer some thoughts on this intriguing topic.

Dreyfus and Taylor (2015) consider the conceptual and the embedded view to be largely incompatible. According to the embedded view, understanding lies in a *preconceptual* engagement with the environment, and that preconceptual form of understanding is circumspection in dealing with equipment. In contrast to that view, McDowell (2013) sees *conceptual* structures in operation already in that context:²⁸

My claim is that capacities that are conceptual, capacities that belong to their possessor's rationality, are operative not only in reflective thought and action but also at the ground-floor level at which there is absorbed coping and acting in flow. (McDowell, 2013, 54)

²⁷ Relevant references that discuss this view and the terminology of embedded-embodied are, e.g., Wheeler (2005) and Dreyfus and Taylor (2015).

²⁸ Part of the McDowell-Dreyfus debate concerns the question whether there really *is* such a contrast. We will here not dwell on the actual exchange of arguments, but only use the debate to anchor the preconceptual-conceptual distinction. For agency considered generally, this contrast is definitely affirmed—for instance, McDowell (2013) makes clear that he focuses on human agency only, and that dogs (his example) do not possess rational agency.

As we have shown, the PS model adopts an embedded view from the framework of reinforcement learning, and we have spelled out a phenomenological interpretation that allows us to describe how absorbed coping can be learnt. At the same time, the PS model also seems to be compatible with McDowell's position once one considers clips or, more generally, certain substructures of the ECM as concepts.

Indeed, we argue that the PS model can illustrate both views and provide a platform where one can model the transition from embodied-embedded towards conceptual-detached aspects of agency, from absorbed coping to detached reflection. In other words, PS can help us to understand the emergence of the conceptual within an embodied-embedded view of preconceptual agency. In the set-up described in Sect. 6.5.1, clips are embodied structures that represent basic encounterings of the agent with its environment. In the language of Dreyfus and Taylor (2015), these clips can be called *preconceptual* insofar as their activation in the context of a deliberation process simply corresponds to a (variational) replay of earlier episodic experiences. This presupposes, plausibly, that these earlier experiences can themselves still be viewed as preconceptual despite their unifying and categorising function.²⁹

We hold that PS offers a way to connect the preconceptual to the conceptual via the emergence of spontaneity and of discriminatory structures that are subject to at least something like rationality constraints.³⁰ This comes about in two steps. First, from a short-time dynamical perspective, the activation of clips in the agent's ECM *decouples* from the direct sensory input of the environment—in particular in deeper clip networks—and takes on a certain life of its own: The activation of a clip triggers a random walk which corresponds to a simple form of deliberation. While firmly integrated in the agent's embedded agency, this decoupling is one of the central functional aspects of the ECM, and it provides the agent with flexibility in reacting to a given perceptual situation. In Sect. 2.6.1 we identified this flexibility as one of the resources of free agency. In line with McDowell (1994), the ECM could therefore be seen as the embodied basis of spontaneity, which is a prerequisite for having conceptual capacities:

But conceptual capacities, capacities that belong to spontaneity, are already at work in experiences themselves, not just in judgements based on them; so experiences can

²⁹ Physically, unification and categorisation or encoding takes place through the interaction of the physical degrees of freedom of the input signal with certain other degrees of freedom of the sensory apparatus of the agent. (A biological example, in the context of visual perception, would be the excitation of cones on the retina that respond to certain optical frequencies in the visual signal.) If such unification were to be called conceptual, one would have to agree that the level of a thermometer's mercury column in direct contact with a heat source would be a concept for the latter's temperature. This, however, completely ignores the normative nature of concepts as elements of an agent's rationality.

³⁰ We are aware that this means stretching, perhaps overstressing the notion of rationality. The agents we consider do not necessarily occupy a Sellarsian space of reasons, which would presuppose that they are able to answer why-questions by giving reasons for their actions. As we argued, agents—PS agents included—need not have a language to begin with. See, however, Sect. 5.7 for a discussion of how PS agents could explain why they did what they did.

intelligibly stand in rational relations to our exercises of the freedom that is implicit in the idea of spontaneity. (McDowell, 1994, 24)

Second, from a (long-time) learning perspective, the decoupling from direct sensory input—while staying within an overall embedded scheme—allows for the development of structures within the ECM that take on further unifying and discriminatory functions (see again Eva et al., 2023). As these discriminatory functions are developed in interaction with an environment that provides feedback via rewards, the long-term learning perspective thus grounds a notion of “getting it right” vs. “getting it wrong” in a way that makes sense to the agent. This captures another aspect of conceptual capacities as elements of an agent’s rationality.

Dreyfus (2013) defends the primacy of the embedded-embodied-contact view against a detached-conceptual-mediational view that McDowell (2013) claims also to reject. A specific target for Dreyfus, also in other writings, is the rejection of representationalism and of the idea of concepts as representing the outside world. Here we also hope to be able to contribute towards a mediating position. In the case of PS agents, we do refer to a representational role of the embodied clip network when we describe it as representing basic encounterings. However, what is represented by the ECM is not the outside world *per se*, but rather the agent’s own embedded agency, as the ECM essentially stores a condensed form of the agent’s history of interactions with the environment. With respect to representations, PS thus also provides room for both views.

In summary, we hold that the role of clips can vary between preconceptual and conceptual, depending on the level of complexity of the agent’s ECM: In the basic model of PS, clips as the units of episodic memory are embodied structures that have a preconceptual role. They are preconceptual insofar as (and as long as) their activation in the course of a deliberation process corresponds to a simple replay of earlier episodic experiences. In more complex ECMs, however, this need no longer be the case. Through the process of learning, structures can develop within ECM that acquire a discriminative and, ultimately, also a conceptual role. This variable role of clips, making room for a conceptual *and* a preconceptual interpretation, thus appears to accord well with the position of Dreyfus and Taylor (2015, 71f.):

We contend that we are able to form conceptual beliefs guided by our surroundings because we live in a preconceptual engagement with these surroundings which involves understanding.

According to this position, concepts are secondary and only emerge within a preconceptual engagement of the agent with her surroundings. That engagement already has its own, preconceptual form of understanding.

For our treatment of agency, questions about the distinction between the preconceptual and the conceptual are interesting, but not central. The PS model offers ways to think about different positions in a neutral and productive way. From a modelling perspective, one can be inspired by ideas from different fields, including phenomenology, philosophy of mind, the cognitive sciences, but also physics and AI. Together, we can move forward to a better understanding of agency at all scales.

References

- Ardón, P., Pairet, È., Lohan, K. S., Ramamoorthy, S., & Petrick, R. P. A. (2020). Affordances in robotic tasks – a survey. arXiv 2004.07400.
- Bahl, S., Mendonca, R., Chen, L., Jain, U., & Pathak, D. (2023). Affordances from human videos as a versatile representation for robotics. *Proceedings of computer vision and pattern recognition*. arXiv 2304.08488.
- Bohg, J., Hausman, K., Sankaran, B., Brock, O., Kragic, D., Schaal, S., & Sukhatme, G. S. (2017). Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33, 1273–1291.
- Braitenberg, V. (1984). *Vehicles*. MIT Press.
- Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, 2, 400.
- Brooks, R. A. (1991). Intelligence without reason. In *Proceedings of the 12th international joint conference on artificial intelligence - Volume 1*, IJCAI'91 (pp. 569–595). Morgan Kaufmann Publishers Inc.
- Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge University Press.
- Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60(23), 685–700.
- Dreyfus, H. L. (1991). *Being-in-the-world: A commentary on Heidegger's Being and Time, Division I*. MIT Press.
- Dreyfus, H. L. (1992). *What computers still can't do: A critique of artificial reason*. The MIT Press.
- Dreyfus, H. L. (1993). Heidegger's critique of the Husserl/Searle account of intentionality. *Social Research*, 60(1), 17–38.
- Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, 171, 1137–1160.
- Dreyfus, H. L. (2013). The myth of the pervasiveness of the mental. In J. K. Schear (Ed.), *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.
- Dreyfus, H. L., & Taylor, C. (2015). *Retrieving Realism*. Harvard University Press.
- Eva, B., Ried, K., Müller, T., & Briegel, H. J. (2023). How a minimal learning agent can infer the existence of unobserved variables in a complex environment. *Minds and Machines*, 33(1), 185–219.
- Ferrero, L. (Ed.). (2022). *The Routledge handbook of philosophy of agency*. Routledge.
- Gallagher, S., & Zahavi, D. (2021). *The Phenomenological Mind* (3rd ed.). Routledge.
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Lawrence Earlbaum.
- Heidegger, M. (1927). *Sein und zeit*. Max Niemeyer. Quotations are from the English translation by John Macquarrie and Edward Robinson, *Being and Time*, Blackwell, Oxford 1962. Page references are to this translation, followed by the page number from the 19th German edition (Tübingen, Max Niemeyer Verlag, 2006) in brackets.
- Heidegger, M. (1962). *Being and time*. Blackwell. Translated from the German *Sein und Zeit* by John Macquarrie and Edward Robinson.
- Heidegger, M. (1982). *The basic problems of phenomenology*. Indiana University Press. Translated by Albert Hofstadter. German original: *Grundbegriffe der Phänomenologie* (Marburg Lectures 1927), *Gesamtausgabe*, Bd. 24, Frankfurt a.M.: Vittorio Klostermann 1975.
- Heidegger, M. (1985). *History of the concept of time: Prolegomena*. Indiana University Press. Translated by Theodore J. Kisiel. German original: *Prolegomena zur Geschichte des Zeitbegriffs* (Marburg Lectures 1925), *Gesamtausgabe*, Bd. 20, Frankfurt a.M.: Vittorio Klostermann 1979.
- Heidegger, M. (1995). *The fundamental concepts of metaphysics: World, finitude, solitude*. Indiana University Press. Translated by William McNeill and Nicholas Walker. German original: *Die Grundbegriffe der Metaphysik: Welt - Endlichkeit - Einsamkeit* (Freiburg Lectures 1929), *Gesamtausgabe*, Bd. 29/30, Frankfurt a.M.: Vittorio Klostermann 1983.

- Heidegger, M. (1996). *Being and time*. State University of New York Press. Translated from the German *Sein und Zeit* by Joan Stambaugh.
- Heidegger, M. (2002). Phenomenological interpretations in connection with Aristotle. An indication of the hermeneutical situation. In J. Van Buren (Ed.), *Supplements: From the earliest essays to being and time and beyond* (pp. 111–146). State University of New York Press. German original: *Phänomenologische Interpretationen zu Aristoteles (Anzeige der hermeneutischen Situation)* (Marburg and Göttingen 1922), *Gesamtausgabe*, Bd. 62, Frankfurt a.M.: Vittorio Klostermann 2005.
- Horton, T. E., Chakraborty, A., & Amant, R. S. (2012). Affordances for robots: a brief survey. *Avant*, 3(2), 70–84.
- Kant, I. (1781). *Kritik der reinen Vernunft*. Hartknoch. English translation by Paul Guyer and Allen W. Wood: *Critique of Pure Reason*, Cambridge: Cambridge University Press 1998. Following custom, references are to the original pagination of the first edition (A) of 1781 and of the second edition (B) of 1787.
- McDowell, J. (2013). The myth of the mind as detached. In J. K. Schear (Ed.), *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.
- McDowell, J. H. (1994). *Mind and world*. Harvard University Press.
- Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Éditions Gallimard. English translation by C. Smith: *Phenomenology of Perception*. Routledge, London 1962. Page references are to this translation.
- Minsky, M. (1967). *Computation: Finite and infinite machines*. Prentice-Hall.
- Norman, D. A. (2013). *The design of everyday things*. Basic Books. Revised and expanded edition.
- Schear, J. K., editor (2013). *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.
- Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Thompson, M. (2004). What is it to wrong someone? A puzzle about justice. In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and value: Themes from the moral philosophy of Joseph Raz* (pp. 333–384). Oxford University Press.
- Trenkwalder, L. M., López-Incera, A., Poulsen Nautrup, H., Flamini, F., & Briegel, H. J. (2023). Automated gadget discovery in the quantum domain. *Machine Learning: Science and Technology*, 4(3), 035043.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2*, 42, 230–265.
- Wheeler, M. (2005). *Reconstructing the cognitive world*. MIT Press.
- Winograd, T., & Flores, F. (1987). *Understanding computers and cognition: A new foundation for design*. Addison Wesley.
- Wittgenstein, L. (1969). *On certainty*. Basil Blackwell. Manuscript dated 1951, translated and edited by Denis Paul and G. E. M. Anscombe.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



Chapter 7

Conclusion



We started this book from a collection of everyday facts that support our firm conviction that we are agents. We act, we make things happen, and what we do makes a difference in the world. Of course.

How does it work, then? How is it that we are agents in our world? At the surface, this looks like a normal scientific question, similar perhaps to the question of how it is that we keep our body temperature at roughly 37°C. But this appearance is misleading. True, scientific investigation will certainly be needed in order to understand our agency and its basis in the biology, physiology, chemistry, and physics of our world. But such an investigation will not be able to make real progress unless the philosophical complexity of the agency question is understood and the question itself is acknowledged as a conceptual *problem of agency*. The problem of agency that we mention in the title of this book is about finding a place for ourselves and other agents, *as agents*, in nature as we understand it through the sciences. How can we conceive of nature, the whole of what there is, such that it provides the wiggle room that agents need in order to make a difference to what happens in nature while they are natural themselves? This reflexive structure creates a challenge: How does agency fit in?

One way to react to this challenge, which might even be the majority view, is agency scepticism: to dismiss the question by rejecting the notion of agency as, basically, a remnant of superstition. If one thinks of nature as a container of lots of stuff, and of agency as operating, as it were, from the outside, then agency must look like a supernatural phenomenon, removed from the kind of understanding that the sciences can offer. We agree that *if* agency had to be understood as a supernatural phenomenon, we would have to reject it; the notion would at least be scientifically sterile. But as we have argued, agency can be understood as a *natural* phenomenon. Agency scepticism may have wide currency, but it is not forced upon us. And it comes at great cost. The fully sceptical route that denies the reality of agency leaves us with the task to abandon our everyday assumptions about our active powers—but even deciding to go down the sceptical route means to exercise the very powers

that would have to be dismissed. Overt agency scepticism is, therefore, practically self-undermining. It is simply not an option to give up on the notion of agency. This leaves the agency sceptic with the equally unpalatable option of trying to retain their everyday assumptions about their active powers while knowing that they are false, which is an exercise in doublethink.

We can do better. But if one rejects agency scepticism and affirms the reality of agency, there is work to do. In our introduction, we presented two versions of the philosophical problem of agency that define the kind of work required to avoid agency scepticism. One worry comes from zooming out of our everyday vision of the world, the other from zooming in. The zooming-out version of the problem notes that whatever it is that agents do, it has to fit into the totality of the worldly drama at a cosmic scale. At that scale, nobody does anything; things simply happen (Nagel, 1986, 114). What remains of agency when we zoom out? Can agency be more than an insignificant twist in the overall course of events?

The mirror image of that worry is triggered by zooming in: At microscopic and smaller scales, everything that happens is governed by biochemical, chemical, and physical processes that are not due to any agent, but which also simply happen. And everything—sand and stars, mice and men—is constituted by microscopic and sub-microscopic entities, whose dynamics is described by the physical laws operating at these smallest scales. What remains of agency when we zoom in? Can agency be more than, at best, a convenient summary term for what really happens down there?

The problem of agency is one of ontology, of what there is, and it arises on any broadly naturalistic metaphysical outlook. Having rejected agency scepticism, an ontological elucidation of agency can start from the basic fact of agency realism: there really are agents. We humans are agents. And animal agency, of which human agency is but one example, has arisen naturally over the course of our long evolutionary history; it is perhaps a surprising, but definitely a wholly worldly affair. It must be that within the range of options offered by nature, some entities have found a way to make use of some of those options for their own ends and thus to become agents. How is this possible, and how did it come about?

Agency can be approached from many different angles, as it lies at the intersection of different fields of study. One highly relevant field is biology. The evolutionary history of how agency has come about is slowly coming into view, elucidating which fitness benefits agency brings for the agent and which mechanisms were involved in making it a reality—witness, for example, the recent accounts by Tomasello (2022) and by Mitchell (2023). Not being biologists, we have little to offer here, but we see our own research as agreeing with and in fact supporting an evolutionary account of agency. What we hope to have contributed in this book is to describe agency and its conceptual and scientific basis from the point of view of three different fields that complement and actually help to ground a biological approach: philosophy, physics, and Artificial Intelligence.

Our key takeaway points are the following.

Agency is an ontologically emergent phenomenon that requires indeterminism at different levels of organisation. Conceptually, it can be clearly separated from the mere causal impact of physical entities and from the high-level, human-specific type of agency that we are familiar with from our own lives. Agency in the sense that interests us is exemplified not just by human beings but also by many animals. Studying our practices of attribution contributes to our explication of agency, which is as follows: An agent is an embodied individual that can settle matters regarding its body, which wants certain things for itself, and whose history is one of learning, developing an individual memory, and adapting to its environment.

Ontological emergence is ubiquitous in nature, and it is instantiated at the purely physical level, as witnessed by the example of the sun. Quantum mechanics provides a view of nature as fundamentally indeterministic, thus offering room for the development of higher-level entities including agents. And agency, in turn, is central for an understanding of the basic conceptual structures of quantum mechanics itself. A quantum state is a tool for an agent that interacts with nature, and calling the agent an observer misses this interactive dimension. This claim is naturally accommodated in the Heisenberg picture of quantum dynamics. In addition, in the development of quantum information science, long-standing philosophical conundra of quantum mechanics have inspired tools for agents to enhance their technological capabilities. The philosophical payoff is that these technologies, which include tuneable randomness, offer a model for the top-down control that agency implies.

Our view of quantum mechanics has much in common with the foundational position of QBism. That position stresses the central systematic role of the agent in quantum mechanics in support of a notion of participatory realism, which agrees with a general view of scientific theories as tools. We go beyond QBism in taking agency itself as something to be understood in detail. Agents build and adapt models in their interaction with nature, which is insufficiently captured by the mere updating of beliefs in a given probability space. We include model building, and agency itself more generally, as a proper target for modelling.

Agency as we explicate it can be modelled in terms of formal structures. Such modelling is a central methodology in physics, and the agency model of Projective Simulation belongs to that tradition. It takes inspiration from indeterminism in physics, reinforcement learning, and core notions of phenomenology. Projective Simulation can be fruitfully applied, e.g., in quantum science and in behavioural biology. Quantum extensions of the model highlight the fundamental and so far unexplored problem of the transparency and explainability of quantum-enhanced Artificial Intelligence.

Projective Simulation can also contribute to the clarification of philosophical aspects of agency, especially as regards phenomenology. Basic notions of existential phenomenology such as circumspection, thrownness and projection can be shown to correspond to basic structures of Projective Simulation. That model, in turn,

helps elucidate and possibly resolve two key lacunae of the Heideggerian approach. Learning and development are central for agency, and especially in basic science, agents encounter and explore yet undetermined entities. Projective Simulation offers a way to model the integration of such entities without presupposing a fixed referential structure. Furthermore, the basic role of randomness and projection anchors the notion of existential possibility in a way that bridges the gap between phenomenology and physics.

*

Modelling offers understanding. In our view, agency modelling and the understanding that comes with it should be especially helpful with respect to major contemporary challenges that we wish to at least mention in closing.

On the one hand, animal agency is a fact. We can understand many animals as agents of various degrees of sophistication. This offers a way to make sense of their behaviour that acknowledges an internal structure of individuality, historicity, and memory. Animal agents are learners and settlers of matters concerning their bodies. This perspective offers insights for the study of animal behaviour, both at the collective and at the individual level. It also forces us to reflect on our interaction with animals as fellow agents.

On the other hand, while artificial agency is not yet a fact, it may soon be. Advances in Artificial Intelligence in a science context are leading towards ever more autonomous science assistant systems, and efforts to develop Artificial General Intelligence may in fact target the development of artificial agency. The societal implications of this ongoing technological transformation are far reaching and provide many challenges of a practical, but also of a conceptual nature. For both dimensions, it will be important to ascertain whether and in which way agency can be properly attributed to a novel technological system. In a science context, the question will be who has *done* something—who has performed an experiment, who has had an insight? This raises the question of the possible role of artificial systems as authors of discoveries, results, or publications, and it forces us to reconsider what an experiment really is.

In a more general context, whether a system is an agent has important implications for its status as the source of an event or the author of a product, and also for questions of responsibility. In our view, it is important to study the agency aspects of newly developed systems as a necessary step towards discussing their “ethics”. Ethics has no application where agency is lacking. The point is fundamental: If a system cannot act, then it cannot *act badly* either—even though its behaviour can be good or bad *for us*. Too often, ethics questions are brought forward and discussed as if they were addressed at the systems under question, while they should properly be addressed at their developers and users. Focusing on the agency of future technological artefacts will be helpful to avoid delegating responsibility to where it can have no foothold in reality.

Even if artificial systems that act become a reality, they will be mere agents, not moral agents. It remains our responsibility to ensure that they benefit human society.

References

- Mitchell, K. J. (2023). *Free agents. How evolution gave us free will*. Princeton University Press.
- Nagel, T. (1986). *The view from nowhere*. Oxford University Press.
- Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.

Open Access This chapter is licensed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>), which permits use, sharing, adaptation, distribution and reproduction in any medium or format, as long as you give appropriate credit to the original author(s) and the source, provide a link to the Creative Commons license and indicate if changes were made.

The images or other third party material in this chapter are included in the chapter's Creative Commons license, unless indicated otherwise in a credit line to the material. If material is not included in the chapter's Creative Commons license and your intended use is not permitted by statutory regulation or exceeds the permitted use, you will need to obtain permission directly from the copyright holder.



References

- Abellán, C., Acín, A., Alarcón, A., & The BIG Bell Test Collaboration. (2018). Challenging local realism with human choices. *Nature*, 557, 212–216.
- Adam, C., & Tannery, P. (1904). *Oeuvres de Descartes*. J. Vrin.
- Adelberger, E. G., García, A., Robertson, R. G. H., Snover, K. A., Balantekin, A. B., Heeger, K., Ramsey-Musolf, M. J., Bemmerer, D., Junghans, A., Bertulani, C. A., Chen, J.-W., Costantini, H., Prati, P., Couder, M., Uberseder, E., Wiescher, M., Cyburt, R., Davids, B., Freedman, S. J., Gai, M., Gazit, D., Gialanella, L., Imbriani, G., Greife, U., Hass, M., Haxton, W. C., Itahashi, T., Kubodera, K., Langanke, K., Leitner, D., Leitner, M., Vetter, P., Winslow, L., Marcucci, L. E., Motobayashi, T., Mukhamedzhanov, A., Tribble, R. E., Nollett, K. M., Nunes, F. M., Park, T.-S., Parker, P. D., Schiavilla, R., Simpson, E. C., Spitaleri, C., Strieder, F., Trautvetter, H.-P., Suemmerer, K., & Typel, S. (2011). Solar fusion cross sections. II. The *pp* chain and CNO cycles. *Reviews of Modern Physics*, 83, 195–245.
- Agresti, I., Poderini, D., Guerini, L., Mancusi, M., Carvachol, G., Aolita, L., Cavalcanti, D., Chaves, R., & Sciarrino, F. (2020). Experimental device-independent certified randomness generation with an instrumental causal structure. *Communications Physics*, 3, 10.
- Alberts, B., Johnson, A., Lewis, J., Morgan, D., Raff, M., Roberts, K., & Walter, P. (2015). *Molecular biology of the cell* (6th ed.). Garland Science.
- Alicki, R., & Lendi, K. (2007). *Quantum dynamical semigroups and applications*. Lecture Notes in Physics (Vol. 717). Springer.
- Allen, L., Beijersbergen, M. W., Spreeuw, R. J. C., & Woerdman, J. P. (1992). Orbital angular momentum of light and the transformation of Laguerre-Gaussian laser modes. *Physical Review A*, 45, 8185–8189.
- Andersen, U. L., Neergaard-Nielsen, J. S., van Loock, P., & Furusawa, A. (2015). Hybrid discrete- and continuous-variable quantum information. *Nature Physics*, 11(9), 713–719.
- Anderson, P. W. (1972). More is different. Broken symmetry and the nature of the hierarchical structure of science. *Science*, 177(4047), 393–396.
- Anscombe, G. E. M. (1957). *Intention*. Blackwell. 2nd edition 1963, reprint Harvard University Press, Cambridge, MA, 2000.
- Anscombe, G. E. M. (1971). *Causality and determination*. Cambridge University Press. Page references are to the reprint in her *Collected Philosophical Papers*, Vol. 2, Oxford 1981, pp 133–147.
- Appleby, M., Flammia, S., McConnell, G., & Yard, J. (2017). SICs and algebraic number theory. *Foundations of Physics*, 47(8), 1042–1059.
- Ardón, P., Pairet, È., Lohan, K. S., Ramamoorthy, S., & Petrick, R. P. A. (2020). Affordances in robotic tasks – a survey. arXiv 2004.07400.

- Arunachalam, S., & de Wolf, R. (2017). Guest column: A survey of quantum learning theory. *SIGACT News*, 48(2), 41–67.
- Aslam, N., Zhou, H., Urbach, E. K., Turner, M. J., Walsworth, R. L., Lukin, M. D., & Park, H. (2023). Quantum sensors for biomedical applications. *Nature Reviews Physics*, 5(3), 157–169.
- Aspect, A., Dalibard, J., & Roger, G. (1982). Experimental test of Bell's inequalities using time-varying analyzers. *Physical Review Letters*, 49, 1804–1807.
- Atkinson, D., & Peijnenburg, J. (2022). How certain is Heisenberg's uncertainty principle? *HOPOS: The Journal of the International Society for the History of Philosophy of Science*, 12(1), 1–21.
- Auletta, G., Fortunato, M., & Parisi, G. (2009). *Quantum mechanics*. Cambridge University Press.
- Austin, J. L. (1956). A plea for excuses. *Proceedings of the Aristotelian Society*, 57, 1–30.
- Austin, J. L. (1962). *Sense and sensibilia*. Oxford University Press.
- Babazadeh, A., Erhard, M., Wang, F., Malik, M., Nouroozi, R., Krenn, M., & Zeilinger, A. (2017). High-dimensional single-photon quantum gates: Concepts and experiments. *Physical Review Letters*, 119, 180510.
- Bacon, F. (1620). *Novum organum*. John Bill. English translation as part of *The Oxford Francis Bacon*, Volume XII: *The instauratio magna: Part II. Novum organum*, edited by G. Rees and M. Wakely. Oxford: Oxford University Press 2007.
- Bagan, E., Bergou, J. A., Cottrell, S. S., & Hillery, M. (2016). Relations between coherence and path information. *Physical Review Letters*, 116, 160406.
- Bagan, E., Calsamiglia, J., Bergou, J. A., & Hillery, M. (2018). A generalized wave-particle duality relation for finite groups. *Journal of Physics A: Mathematical and Theoretical*, 51(41), 414015.
- Bahcall, J. N., & May, R. M. (1969). The rate of the proton-proton reaction and some related reactions. *The Astrophysical Journal*, 155, 501.
- Bahl, S., Mendonca, R., Chen, L., Jain, U., & Pathak, D. (2023). Affordances from human videos as a versatile representation for robotics. *Proceedings of computer vision and pattern recognition*. arXiv 2304.08488.
- Ballentine, L. E. (2015). *Quantum mechanics. A modern development* (2nd ed.). World Scientific.
- Barnett, S. M. (2009). *Quantum information*. Oxford Master Series in Atomic, Optical, and Laser Physics. Oxford University Press.
- Barreiro, J. T., Müller, M., Schindler, P., Nigg, D., Monz, T., Chwalla, M., Hennrich, M., Roos, C. F., Zoller, P., & Blatt, R. (2011). An open-system quantum simulator with trapped ions. *Nature*, 470(7335), 486–491.
- Barto, A. G. (2013). Intrinsic motivation and reinforcement learning. In G. Baldassarre, & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems* (pp. 17–47). Springer.
- Bauer, P., Stevens, B., & Hazeleger, W. (2021). A digital twin of Earth for the green transition. *Nature Climate Change*, 11, 80–83.
- Bäckström, S. and Gustafsson, M. (2017). Skill, drill, and intelligent performance: Ryle and intellectualism. *Journal for the History of Analytical Philosophy*, 5(5):40–55.
- Bell, J. S. (1964). On the Einstein Podolsky Rosen paradox. *Physics*, 1, 195–200. Reprinted as Ch. 2 of Bell's *Speakable and Unsayable in Quantum Mechanics* (1987).
- Bell, J. S. (1966). On the problem of hidden variables in quantum mechanics. *Reviews of Modern Physics*, 38, 447–452. Reprinted as Ch. 1 of Bell's *Speakable and Unsayable in Quantum Mechanics* (1987).
- Bell, J. S. (1987). *Speakable and unsayable in quantum mechanics*. Cambridge University Press.
- Bell, J. S. (1990). Against 'measurement'. *Physics World*, 3, 33–40.
- Bellman, R. E. (1957). *Dynamic programming*. Princeton University Press.
- Belnap, N., Müller, T., & Placek, T. (2022). *Branching space-times. theory and applications*. Oxford studies in philosophy of science. Oxford University Press.
- Belnap, N., Perloff, M., & Xu, M. (2001). *Facing the future: Agents and choices in our indeterminist world*. Oxford University Press.
- Bengtsson, I. (2020). SICs: Some explanations. *Foundations of Physics*, 50(12), 1794–1808.

- Bengtsson, I., & Życzkowski, K. (2017). *Geometry of quantum states: An introduction to quantum entanglement* (2nd ed.). Cambridge University Press.
- Bennett, C. H., Brassard, G., Crépau, C., Jozsa, R., Peres, A., & Wootters, W. K. (1993). Teleporting an unknown quantum state via dual classical and Einstein-Podolsky-Rosen channels. *Physical Review Letters*, 70, 1895–1899.
- Bennett, C. H., Brassard, G., Popescu, S., Schumacher, B., Smolin, J. A., & Wootters, W. K. (1996). Purification of noisy entanglement and faithful teleportation via noisy channels. *Physical Review Letters*, 76, 722–725.
- Bennett, C. H., & Wiesner, S. J. (1992). Communication via one- and two-particle operators on Einstein-Podolsky-Rosen states. *Physical Review Letters*, 69, 2881–2884.
- Berryman, S. (2022). Ancient atomism. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Winter 2022 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2022/entries/atomism-ancient/>
- Biamonte, J., Wittek, P., Pancotti, N., Rebentrost, P., & Lloyd, S. (2017). Quantum machine learning. *Nature*, 549, 195–202.
- Bierhorst, P., Knill, E., Glancy, S., Zhang, Y., Mink, A., Jordan, S., Rommal, A., Liu, Y.-K., Christensen, B., Nam, S. W., et al. (2018). Experimentally generated randomness certified by the impossibility of superluminal signals. *Nature*, 556(7700), 223–226.
- Bjerland, Ø. F. (2015). Projective simulation compared to reinforcement learning. Master's thesis, Department of Informatics, University of Bergen. <https://hdl.handle.net/1956/10391>
- Black, M. (1946). *Critical thinking. An introduction to logic and scientific method*. Prentice Hall.
- Bluvstein, D., Evered, S. J., Geim, A. A., Li, S. H., Zhou, H., Manovitz, T., Ebadi, S., Cain, M., Kalinowski, M., Hangleiter, D., Bonilla Ataides, J. P., Maskara, N., Cong, I., Gao, X., Sales Rodriguez, P., Karolyshyn, T., Semeghini, G., Gullans, M. J., Greiner, M., Vuletić, V., & Lukin, M. D. (2024). Logical quantum processor based on reconfigurable atom arrays. *Nature*, 626(7997), 58–65.
- Bluvstein, D., Levine, H., Semeghini, G., Wang, T. T., Ebadi, S., Kalinowski, M., Keesling, A., Maskara, N., Pichler, H., Greiner, M., Vuletić, V., & Lukin, M. D. (2022). A quantum processor based on coherent transport of entangled atom arrays. *Nature*, 604(7906), 451–456.
- Bohg, J., Hausman, K., Sankaran, B., Brock, O., Kragic, D., Schaal, S., & Sukhatme, G. S. (2017). Interactive perception: Leveraging action in perception and perception in action. *IEEE Transactions on Robotics*, 33, 1273–1291.
- Bohm, D. (1952). A suggested interpretation of the quantum theory in terms of “hidden” variables. I. *Physical Review*, 85, 166–179.
- Bohm, D., & Hiley, B. J. (1993). *The undivided universe: An ontological interpretation of quantum theory*. Routledge.
- Bohr, N. (1934). *Atomic theory and the description of nature*. Cambridge University Press.
- Borexino Collaboration. (2018). Comprehensive measurement of pp-chain solar neutrinos. *Nature*, 562, 505–510.
- Borges, J. L. (1946). Del rigor en la ciencia. In *El Hacedor*, 1961. Translated by Andrew Hurley, Borges, *Collected Fictions*, Penguin 1998, as “On Exactitude in Science”. Newer translated by Diego Doval as “On rigor in science”. <https://medium.com>, 2017.
- Boyajian, W. L., Clausen, J., Trenkwalder, L. M., Dunjko, V., & Briegel, H. J. (2020). On the convergence of projective-simulation-based reinforcement learning in Markov decision processes. *Quantum Machine Intelligence*, 2(2), 13.
- Braitenberg, V. (1984). *Vehicles*. MIT Press.
- Bratman, M. E. (1987). *Intention, plans, and practical reason*. Harvard University Press.
- Bravyi, S., Cross, A. W., Gambetta, J. M., Maslov, D., Rall, P., & Yoder, T. J. (2024). High-threshold and low-overhead fault-tolerant quantum memory. *Nature*, 627, 778–782.
- Brentano, F. (1874). *Psychologie vom empirischen Standpunkte*. Dunker and Humbolt. References are to the English translation, *Psychology from an Empirical Standpoint*, Routledge Classics ed., with a foreword by Tim Crane, London 2015.

- Briegel, H. J. (2001). Principles of entanglement purification. In D. Bouwmeester, A. Ekert, A. & Zeilinger (Eds.), *The physics of quantum information: Quantum cryptography, quantum teleportation, quantum computation* (pp. 261–268). Springer.
- Briegel, H. J. (2009). Cluster states. In D. Greenberger, K. Hentschel, & F. Weinert (Eds.), *Compendium of quantum physics* (pp. 96–105). Springer.
- Briegel, H. J. (2012). On creative machines and the physical origins of freedom. *Scientific Reports*, 2, 522.
- Briegel, H. J., & De las Cuevas, G. (2012). Projective simulation for artificial intelligence. *Scientific Reports*, 2, 400.
- Briegel, H.-J., Dür, W., Cirac, J. I., & Zoller, P. (1998). Quantum repeaters: The role of imperfect local operations in quantum communication. *Physical Review Letters*, 81, 5932–5935.
- Briegel, H. J., & Müller, T. (2015). A chance for attributable agency. *Minds and Machines*, 25, 261–279.
- Briegel, H. J., & Raussendorf, R. (2001). Persistent entanglement in arrays of interacting particles. *Physical Review Letters*, 86, 910–913.
- Briegel, K. D., von Grafenstein, N. R., Draeger, J. C., Blümmler, P., Allert, R. D., & Bucher, D. B. (2025). Optical widefield nuclear magnetic resonance microscopy. *Nature Communications*, 16(1), 1281.
- Brooks, R. A. (1991). Intelligence without reason. In *Proceedings of the 12th International Joint Conference on Artificial Intelligence - Volume 1, IJCAI'91* (pp. 569–595). Morgan Kaufmann Publishers Inc.
- Busch, P., Lahti, P. J., & Mittelstaedt, P. (1996). *The quantum theory of measurement* (Vol. 2, 2nd revised ed.). Lecture notes in physics monographs. Springer.
- Bustamante, C. J., Chemla, Y. R., Liu, S., & Wang, M. D. (2021). Optical tweezers in single-molecule biophysics. *Nature Reviews Methods Primers*, 1(1), 25.
- Butterfield, J. (2005). Determinism and indeterminism. In *Routledge encyclopedia of philosophy* (Vol. 3). Routledge.
- Butterfield, J. (2011a). Emergence, reduction and supervenience: A varied landscape. *Foundations of Physics*, 41(6), 920–959.
- Butterfield, J. (2011b). Less is different: Emergence and reduction reconciled. *Foundations of Physics*, 41(6), 1065–1135.
- Cabello, A., Estebaranz, J., & García-Alcaine, G. (1996). Bell-Kochen-Specker theorem: A proof with 18 vectors. *Physics Letters A*, 212(4), 183–187.
- Calsamiglia, J., Hartmann, L., Dür, W., & Briegel, H.-J. (2005). Spin gases: Quantum entanglement driven by classical kinematics. *Physical Review Letters*, 95, 180502.
- Candlish, S., & Wrisley, G. (2019). Private language. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2019 edition. <https://plato.stanford.edu/archives/fall2019/entries/private-language/>
- Carlson, J., Riska, D. O., Schiavilla, R., & Wiringa, R. B. (1991). Weak proton capture reactions on ^1H and ^3He and tritium β decay. *Physical Review C*, 44(2), 619–625.
- Carneiro, R. L. (2000). The transition from quantity to quality: A neglected causal mechanism in accounting for social evolution. *Proceedings of the National Academy of Sciences*, 97(23), 12926–12931.
- Cartwright, N. (1999). *The dappled world. A study of the boundaries of science*. Cambridge University Press.
- Cerezo, M., Arrasmith, A., Babbush, R., Benjamin, S. C., Endo, S., Fujii, K., McClean, J. R., Mitarai, K., Yuan, X., Cincio, L., & Coles, P. J. (2021). Variational quantum algorithms. *Nature Reviews Physics*, 3(9), 625–644.
- Cerf, N. J., Bourennane, M., Karlsson, A., & Gisin, N. (2002). Security of quantum key distribution using d -level systems. *Physical Review Letters*, 88, 127902.
- Chen, S. Y.-C., Yang, C.-H. H., Qi, J., Chen, P.-Y., Ma, X., & Goan, H.-S. (2020). Variational quantum circuits for deep reinforcement learning. *IEEE Access*, 8, 141007–141024.
- Chi, X., Li, Y., & Qiu, X. (2020). V(D)J recombination, somatic hypermutation and class switch recombination of immunoglobulins: Mechanism and regulation. *Immunology*, 160(3), 233–247.

- Chiribella, G., D'Ariano, G. M., & Perinotti, P. (2009). Theoretical framework for quantum networks. *Physical Review A*, 80, 022339.
- Christensen, N., & Meyer, R. (2022). Parameter estimation with gravitational waves. *Reviews of Modern Physics*, 94, 025001.
- Clancey, W. J. (1997). *Situated cognition: On human knowledge and computer representations*. Cambridge University Press.
- Clarke, R., Capes, J., & Swenson, P. (2021). Incompatibilist (nondeterministic) theories of free will. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Fall 2021 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2021/entries/incompatibilism-theories/>
- Clements, W. R., Humphreys, P. C., Metcalf, B. J., Kolthammer, W. S., & Walmsley, I. A. (2016). Optimal design for universal multiport interferometers. *Optica*, 3(12), 1460–1465.
- Cloeren, H. J. (1984). Ockham's razor. In J. Ritter & K. und Gründer (Eds.), *Historisches Wörterbuch der Philosophie*, Bd. 6. Schwabe Verlag, Sp. 1094–1095.
- Cohen-Tannoudji, C., Diu, B., & Laloë, F. (2019). *Quantum mechanics: Volume I: Basic concepts, tools, and applications* (2nd ed.). Wiley-VCH.
- Commins, E. D. (2012). Electron spin and its history. *Annual Review of Nuclear and Particle Science*, 62(1), 133–157.
- Cottingham, J., Stoothoff, R., & Murdoch, D. (1984). *The philosophical writings of Descartes*. Cambridge University Press.
- Cronin, A. D., Schmiedmayer, J., & Pritchard, D. E. (2009). Optics and interferometry with atoms and molecules. *Reviews of Modern Physics*, 81, 1051–1129.
- Cybert, R. H., Fields, B. D., Olive, K. A., & Yeh, T.-H. (2016). Big bang nucleosynthesis: Present status. *Reviews of Modern Physics*, 88, 015004.
- Davidson, D. (1963). Actions, reasons, and causes. *The Journal of Philosophy*, 60(23), 685–700.
- Davidson, D. (1978). Intending. *Philosophy of History and Action*, 11, 41–60.
- De Finetti, B. (2017). *Theory of probability. A critical introductory treatment* (Combined edition.). Wiley Two volumes in one book. Translated from the 1970 Italian original, *Teoria Delle Probabilita*, by Antonio Machí and Adrian Smith.
- De Regt, H. W. (2017). *Understanding scientific understanding*. Oxford Studies in Philosophy of Science. Oxford University Press.
- DeBroda, J. B., Fuchs, C. A., Pienaar, J. L., & Stacey, B. C. (2021). Born's rule as a quantum extension of Bayesian coherence. *Physical Review A*, 104:022207.
- DeBroda, J. B., Fuchs, C. A., & Schack, R. (2020). Respecting one's fellow: QBism's analysis of Wigner's friend. *Foundations of Physics*, 50, 1859–1874.
- DeBroda, J. B., Fuchs, C. A., & Schack, R. (2024). QBism's account of quantum dynamics and decoherence. *Physical Review A*, 110, 052205.
- DeBroda, J. B., & Love, P. J. (2022). Quantum and classical Bayesian agents. *Quantum*, 6, 713.
- DeBroda, J. B., & Stacey, B. C. (2018). FAQBism. arXiv 1810.13401.
- Del Colle, R. (2011). Miracles in Christianity. In G. H. Twelftree (Ed.) *The Cambridge companion to miracles* (pp. 235–253). Cambridge University Press.
- Delfosse, N., & Nickerson, N. H. (2021). Almost-linear time decoding algorithm for topological codes. *Quantum*, 5, 595.
- Dennett, D. C. (1984). *The intentional stance*. MIT Press.
- Dennett, D. C. (1991). *Consciousness explained*. Back Bay Books.
- Dennett, D. C. (2003). *Freedom evolves*. Vinking Books.
- Dennis, E., Kitaev, A., Landahl, A., & Preskill, J. (2002). Topological quantum memory. *Journal of Mathematical Physics*, 43(9), 4452–4505.
- Deutsch, D., Ekert, A., Jozsa, R., Macchiavello, C., Popescu, S., & Sanpera, A. (1996). Quantum privacy amplification and the security of quantum cryptography over noisy channels. *Physical Review Letters*, 77, 2818–2821.
- Diaconis, P., Holmes, S., & Montgomery, R. (2007). Dynamical bias in the coin toss. *SIAM Review*, 49(2), 211–235.
- Dieks, D. (1982). Communication by EPR devices. *Physics Letters A*, 92(6), 271–272.

- Dieks, D. (2022). Perspectival quantum realism. *Foundations of Physics*, 52, 95.
- Diestel, R. (2017). *Graph theory*. Graduate texts in mathematics (vol. 173). Springer.
- DiNardoDiNardo, J. (2016). Natural experiments and quasi-natural experiments. In M. Vernengo, E.P. Caldentey & B.J.J. Rosser (Eds.), *The new palgrave dictionary of economics* (pp. 1–12). Palgrave Macmillan.
- Douguet, N., & Bartschat, K. (2018). Dynamics of tunneling ionization using Bohmian mechanics. *Physical Review A*, 97, 013402.
- Dowty, D. R. (1977). Toward a semantic analysis of verb aspect and the English ‘imperfective’ progressive. *Linguistics and Philosophy*, 1, 45–77.
- Dretske, F. (1994). If you can’t make one, you don’t know how it works. *Midwest Studies in Philosophy*, 19, 468–482.
- Dreyfus, H. L. (1991). *Being-in-the-world: A commentary on Heidegger’s Being and Time, Division I*. MIT Press.
- Dreyfus, H. L. (1992). *What computers still can’t do: A critique of artificial reason*. The MIT Press.
- Dreyfus, H. L. (1993). Heidegger’s critique of the Husserl/Searle account of intentionality. *Social Research*, 60(1), 17–38.
- Dreyfus, H. L. (2007). Why Heideggerian AI failed and how fixing it would require making it more Heideggerian. *Artificial Intelligence*, 171, 1137–1160.
- Dreyfus, H. L. (2013). The myth of the pervasiveness of the mental. In J. K. Schear (Ed.), *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.
- Dreyfus, H. L., & Taylor, C. (2015). *Retrieving Realism*. Harvard University Press.
- Driesch, H. (1914). *The history and theory of vitalism*. Macmillan.
- Ducheyne, S. (2014). Newton on action at a distance. *Journal of the History of Philosophy*, 52(4), 675–701.
- Dunjko, V., & Briegel, H. J. (2015). Quantum mixing of Markov chains for special distributions. *New Journal of Physics*, 17(7), 073004.
- Dunjko, V., & Briegel, H. J. (2018). Machine learning & artificial intelligence in the quantum domain: A review of recent progress. *Reports on Progress in Physics*, 81, 074001.
- Dunjko, V., Taylor, J. M., & Briegel, H. J. (2016). Quantum-enhanced machine learning. *Physical Review Letters*, 117, 130501.
- Dunjko, V., Taylor, J. M., & Briegel, H. J. (2017). Advances in quantum reinforcement learning. In *2017 IEEE International Conference on Systems, Man, and Cybernetics (SMC)* (pp. 282–287).
- Dupré, J. (1995). *The disorder of things. Metaphysical foundations of the disunity of science*. Harvard University Press.
- Dupré, J. (2021). Causally powerful processes. *Synthese*, 199, 10667–10683.
- Dürr, D., Goldstein, S., Tumulka, R., & Zanghì, N. (2004). Bohmian mechanics and quantum field theory. *Physical Review Letters*, 93, 090402.
- Dürr, S. (2001). Quantitative wave-particle duality in multibeam interferometers. *Physical Review A*, 64, 042113.
- Earman, J. (1986). *A primer on determinism*. D. Reidel.
- Earman, J. (2019). Quantum Bayesianism assessed. *The Monist*, 102(4):403–423.
- Echeñique, J. (2012). *Aristotle’s ethics and moral responsibility*. Cambridge University Press.
- Edmonds, J. (1965). Paths, trees, and flowers. *Canadian Journal of Mathematics*, 17, 449–467.
- Einstein, A., & Ehrenfest, P. (1922). Quantentheoretische Bemerkungen zum Experiment von Stern und Gerlach. *Zeitschrift für Physik*, 11, 31–34.
- Einstein, A., Podolsky, B., & Rosen, N. (1935). Can quantum-mechanical description of physical reality be considered complete? *Physical Review*, 47(10), 777–780.
- Ekert, A. (1991). Quantum cryptography based on Bell’s theorem. *Physical Review Letters*, 67, 661–663.
- Elliott, T. J., Gu, M., Garner, A. J. P., & Thompson, J. (2022). Quantum adaptive agents with efficient long-term memories. *Physical Review X*, 12, 011007.
- Englert, B.-G. (1996). Fringe visibility and which-way information: An inequality. *Physical Review Letters*, 77, 2154.

- Englert, B.-G. (2006a). *Lectures on quantum mechanics. Volume 1: Basic matters*. World Scientific.
- Englert, B.-G. (2006b). *Lectures on quantum mechanics. Volume 3: Perturbed evolution*. World Scientific.
- Englert, B.-G. (2013). On quantum theory. *European Physical Journal D*, 67, 238.
- Englert, B.-G., Schwinger, J., & Scully, M. O. (1988). Is spin coherence like Humpty-Dumpty? I. Simplified treatment. *Foundations of Physics*, 18, 1045–1056.
- Erhard, A., Poulsen Nautrup, H., Meth, M., Postler, L., Stricker, R., Stadler, M., Negnevitsky, V., Ringbauer, M., Schindler, P., Briegel, H. J., Blatt, R., Friis, N., & Monz, T. (2021). Entangling logical qubits with lattice surgery. *Nature*, 589(7841), 220–224.
- European Union. (2024). Quantum flagship projects. <https://qt.eu/projects/>.
- Eva, B., Ried, K., Müller, T., & Briegel, H. J. (2023). How a minimal learning agent can infer the existence of unobserved variables in a complex environment. *Minds and Machines*, 33(1), 185–219.
- Falkenburg, B., & Morrison, M. (Eds.). (2015). *Why more is different*. Springer.
- Felappi, G. (2023). Propositionalism and questions that do not have correct answers. *Erkenntnis*, 88, 2063–2081.
- Ferrero, L. (Ed.). (2022). *The Routledge handbook of philosophy of agency*. Routledge.
- Feynman, R. P., Leighton, R. B., & Sands, M. I. (1989). *The feynman lectures on physics* (Vol. 3, Commemorative ed.). Addison-Wesley.
- Flamini, F., Hamann, A., Jerbi, S., Trenkwalder, L. M., Poulsen Nautrup, H., & Briegel, H. J. (2020). Photonic architecture for reinforcement learning. *New Journal of Physics*, 22(4), 045002.
- Flamini, F., Krumm, M., Fiderer, L. J., Müller, T., & Briegel, H. J. (2024). Towards interpretable quantum machine learning via single-photon quantum walks. *Quantum Science and Technology*, 9, 045011.
- Forman, P. (1968). The doublet riddle and atomic physics circa 1924. *Isis*, 59, 156–174.
- Fossa, F. (2021). Artificial agency and the game of semantic extension. *Interdisciplinary Science Reviews*, 46(4), 440–457.
- Fowler, A. G., Mariantoni, M., Martinis, J. M., & Cleland, A. N. (2012a). Surface codes: Towards practical large-scale quantum computation. *Physical Review A*, 86, 032324.
- Fowler, A. G., Whiteside, A. C., & Hollenberg, L. C. L. (2012b). Towards practical classical processing for the surface code. *Physical Review Letters*, 108, 180501.
- Franceschetto, G., & Ricou, A. (2024). Demonstration of quantum projective simulation on a single-photon-based quantum computer. *Physical Review A*, 110, 062613.
- Frankfurt, H. G. (1969). Alternate possibilities and moral responsibility. *Journal of Philosophy*, 66(23), 829–839.
- Frauchiger, D., & Renner, R. (2018). Quantum theory cannot consistently describe the use of itself. *Nature Communications*, 9, 3711.
- Frege, G. (1893). *Grundgesetze der Arithmetik, Band I*. Verlag Hermann Pohle. English version, translated and edited by Philip A. Ebert and Marcus Rossberg, as *Gottlob Frege: Basic Laws of Arithmetic*, Oxford: Oxford University Press 2013.
- French, S. (2019). Identity and individuality in quantum theory. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy* (Winter 2019 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2019/entries/qt-idind/>
- French, S. (2023a). *A phenomenological approach to quantum mechanics. Cutting the chain of correlations*. Oxford University Press.
- French, S. (2023b). Putting some flesh on the participant in participatory realism. In P. Berghofer & H. A. Wilsche (Eds.), *Phenomenology and QBism: New approaches to quantum mechanics* (pp. 268–289). Routledge.
- Friedrich, B. (2023). A century ago the Stern-Gerlach experiment ruled unequivocally in favor of quantum mechanics. *Israel Journal of Chemistry*, 63, e202300047.
- Friedrich, B., & Herschbach, D. (2003). Stern and Gerlach: How a bad cigar helped reorient atomic physics. *Physics Today*, 56, 53–59.

- Frigg, R., & Hartmann, S. (2025). Models in science. In E. N. Zalta (Ed.), *The Stanford encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2025 edition. <https://plato.stanford.edu/archives/sum2025/entries/models-science/>
- Friis, N., Vitagliano, G., Malik, M., & Huber, M. (2019). Entanglement certification from theory to experiment. *Nature Reviews Physics*, 1(1), 72–87.
- Frisch, R., & Segré, E. (1933). Über die Einstellung der Richtungsquantelung. II. *Zeitschrift für Physik*, 80, 610–616.
- Fuchs, C. A. (2010). QBism, the perimeter of quantum Bayesianism. arXiv 1003.5209.
- Fuchs, C. A. (2017a). Notwithstanding Bohr, the reasons for QBism. *Mind and Matter*, 15(2), 245–300.
- Fuchs, C. A. (2017b). On participatory realism. In I. T. Durham & D. Rickles (Eds.), *Information and interaction: Eddington, Wheeler, and the limits of knowledge* (pp. 113–134). Springer.
- Fuchs, C. A., Hoang, M. C., & Stacey, B. C. (2017). The SIC question: History and state of play. *Axioms*, 6(3), 21.
- Fuchs, C. A., Mermin, N. D., & Schack, R. (2014). An introduction to QBism with an application to the locality of quantum mechanics. *American Journal of Physics*, 82(8), 749–754.
- Fuchs, C. A., Olshani, M., & Weiss, M. B. (2022). Quantum mechanics? It's all fun and games until someone loses an *i*. arXiv 2206.15343.
- Fuchs, C. A., & Schack, R. (2013). Quantum-Bayesian coherence. *Reviews of Modern Physics*, 85, 1693–1715.
- Fuchs, C. A., & Schack, R. (2015). QBism and the Greeks: Why a quantum state does not represent an element of physical reality. *Physica Scripta*, 90(1), 015104.
- Gachechiladze, M., Budroni, C., & Gühne, O. (2016). Extreme violation of local realism in quantum hypergraph states. *Physical Review Letters*, 116, 070401.
- Galindo, A., & Pascual, P. (1990). *Quantum mechanics I*. Theoretical and mathematical physics. Springer.
- Gallagher, S., & Zahavi, D. (2021). *The Phenomenological Mind* (3rd ed.). Routledge.
- Gallow, J. D. (2022). The metaphysics of causation. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2022 edition. <https://plato.stanford.edu/archives/sum2022/entries/causation-metaphysics/>
- Gaspar, J. M., Bergerault, A., Apostolou, V., & Ricou, A. (2024). Entanglement-enhanced quantum reinforcement learning: an application using single-photons. In *2024 IEEE international conference on quantum computing and engineering (QCE)* (Vo. 02, pp. 329–334).
- Gentile, A. A., Flynn, B., Knauer, S., Wiebe, N., Paesani, S., Granade, C. E., Rarity, J. G., Santagati, R., & Laing, A. (2021). Learning models of quantum systems from experiments. *Nature Physics*, 17(7), 837–843.
- Gerlach, W., & Stern, O. (1922). Der experimentelle Nachweis der Richtungsquantelung im Magnetfeld. *Zeitschrift für Physik*, 9, 349–352.
- Gettier, E. L. (1963). Is justified true belief knowledge? *Analysis*, 23, 121–123.
- Gühne, O., & Tóth, G. (2009). Entanglement detection. *Physics Reports*, 474(1), 1–75.
- Gibson, J. J. (1986). *The ecological approach to visual perception*. Lawrence Earlbaum.
- Giustina, M., Versteegh, M. A. M., Wengerowsky, S., Handsteiner, J., Hochrainer, A., Phelan, K., Steinlechner, F., Kofler, J., Larsson, J.-A., Abellán, C., Amaya, W., Pruneri, V., Mitchell, M. W., Beyer, J., Gerrits, T., Lita, A. E., Shalm, L. K., Nam, S. W., Scheidl, T., Ursin, R., Wittmann, B., & Zeilinger, A. (2015). Significant loophole-free test of Bell's theorem with entangled photons. *Physical Review Letters*, 115, 250401.
- Glauber, R. J. (1986). Amplifiers, attenuators, and Schrödinger's cat. *Annals of the New York Academy of Sciences*, 480(1), 336–372.
- Gleason, A. M. (1957). Measures on the closed subspaces of a Hilbert space. *Indiana University Mathematics Journal*, 6(4), 885–893.
- Glesener, G. (2020). Adaptive measurement-based quantum computation using projective simulation in the presence of graph state symmetries. Master's thesis, University of Innsbruck.
- Glick, D. (2021). QBism and the limits of scientific realism. *European Journal for Philosophy of Science*, 11, 53.

- Goffman, E. (1956). *The presentation of self in everyday life*. University of Edinburgh Social Sciences Research Centre.
- Goldstein, H., Poole, C. P., & Safko, J. L. (2002). *Classical mechanics* (3rd ed.). Addison-Wesley.
- Goldstein, J. A. (2014). Anderson, P.W., “More is different. Broken symmetry and the nature of the hierarchical structure of science”, reprinted with an introduction. *Emergence: Complexity and Organization*, 16(3), 117–134.
- Gottesman, D. (1997). Stabilizer codes and quantum error correction. arXiv quant-ph/9705052.
- Gottesman, D., Kitaev, A., & Preskill, J. (2001). Encoding a qubit in an oscillator. *Physical Review A*, 64, 012310.
- Goudsmit, S. A. (1971). De ontdekking van de elektronenrotatie. *Nederlands Tijdschrift voor Natuurkunde*, 37(16), 386–392. English translation as “The discovery of the electron spin” (by J.H. van der Waals) in *Foundations of Modern EPR*, ed. G.R. Eaton, S.S. Eaton, and K.M. Salikhov, World Scientific, Singapore, 1998, pp. 1–12.
- Granade, C., Combes, J., & Cory, D. G. (2016). Practical Bayesian tomography. *New Journal of Physics*, 18(3), 033024.
- Grassl, M. (2023). New quantum codes from CSS codes. *Quantum Information Processing*, 22(1), 86.
- Grassl, M., Klappenecker, A., & Rötteler, M. (2002). Graphs, quadratic forms, and quantum codes. *Proceedings of the international symposium on information theory, Lausanne, June/July 2002*. arXiv quant-ph/0703112.
- Gray, K., & Wegner, D. (2010). Blaming God for our pain: Human suffering and the divine mind. *Personality and Social Psychology Review*, 14, 7–16.
- Greenberger, D. M., Horne, M. A., Shimony, A., & Zeilinger, A. (1990). Bell’s theorem without inequalities. *American Journal of Physics*, 58, 1131–1143.
- Greenberger, D. M., & Yasin, A. (1988). Simultaneous wave and particle knowledge in a neutron interferometer. *Physics Letters A*, 128(8), 391–394.
- Gröblacher, S., Paterek, T., Kaltenbaek, R., Brukner, Č., Żukowski, M., Aspelmeyer, M., & Zeilinger, A. (2007). An experimental test of non-local realism. *Nature*, 446(7138), 871–875.
- Gross, D., Liu, Y.-K., Flammia, S. T., Becker, S., & Eisert, J. (2010). Quantum state tomography via compressed sensing. *Physical Review Letters*, 105, 150401.
- Gu, M., Weedbrook, C., Perales, A., & Nielsen, M. A. (2009). More really is different. *Physica D: Nonlinear Phenomena*, 238(9–10), 835–839.
- Gühne, O., Tóth, G., Hyllus, P., & Briegel, H. J. (2005). Bell inequalities for graph states. *Physical Review Letters*, 95, 120405.
- Gunning, D., Stefik, M., Choi, J., Miller, T., Stumpf, S., & Yang, G.-Z. (2019). XAI—Explainable artificial intelligence. *Science Robotics*, 4(37), eaay7120.
- Gunning, D., Vorm, E., Wang, J. Y., & Turek, M. (2021). Darpa’s explainable AI (XAI) program: A retrospective. *Applied AI Letters*, 2(4), e61.
- Gustafson, S. J., & Sigal, I. M. (2011). *Mathematical concepts of quantum mechanics*. Universitext. Springer.
- Hacking, I. (1983). *Representing and intervening*. Cambridge University Press.
- Hacking, I. (1990). *The taming of chance*. Cambridge University Press.
- Hafner, V., Loviken, P., Pico Villalpando, A., & Schillaci, G. (2020). Prerequisites for an artificial self. *Frontiers in Neurorobotics*, 14, 5.
- Hangl, S., Dunjko, V., Briegel, H. J., & Piater, J. (2020). Skill learning by autonomous robotic playing using active learning and exploratory behavior composition. *Frontiers in Robotics and AI*, 7, 42.
- Hangl, S., Ugur, E., Szedmak, S., & Piater, J. (2016). Robotic playing for hierarchical complex skill learning. In *Intelligent robots and systems (IROS), 2016 IEEE/RSJ international conference* (pp. 2799–2804).
- Harrington, P. M., Mueller, E. J., & Murch, K. W. (2022). Engineered dissipation for quantum information science. *Nature Reviews Physics*, 4(10), 660–671.
- Hartman, L. (2013). “Que sera sera”: The English roots of a pseudo-Spanish proverb. *Proverbium*, 30, 51–104.

- Hayes, W. B. (2007). Is the outer solar system chaotic? *Nature Physics*, 3, 689–691.
- Head-Gordon, M., & Artacho, E. (2008). Chemistry on the computer. *Physics Today*, 61(4), 58–63.
- Healey, R. (2023). Quantum-Bayesian and pragmatist views of quantum theory. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (winter 2023 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/win2023/entries/quantum-bayesian/>.
- Hegel, G. F. W. (1812). *Wissenschaft der Logik*. Johann Leonhard Schrag. Translated by A. V. Miller as *Hegel's Science of Logic*, London: Allen and Unwin, 1969. References are to this translation.
- Heidegger, M. (1927). *Sein und Zeit*. Max Niemeyer. Quotations are from the English translation by John Macquarrie and Edward Robinson, *Being and Time*, Blackwell, Oxford 1962. Page references are to this translation, followed by the page number from the 19th German edition (Tübingen, Max Niemeyer Verlag, 2006) in brackets.
- Heidegger, M. (1962). *Being and time*. Blackwell. Translated from the German *Sein und Zeit* by John Macquarrie and Edward Robinson.
- Heidegger, M. (1982). *The basic problems of phenomenology*. Indiana University Press. Translated by Albert Hofstadter. German original: *Grundbegriffe der Phänomenologie* (Marburg Lectures 1927), *Gesamtausgabe*, Bd. 24, Frankfurt a.M.: Vittorio Klostermann 1975.
- Heidegger, M. (1985). *History of the concept of time: Prolegomena*. Indiana University Press. Translated by Theodore J. Kisiel. German original: *Prolegomena zur Geschichte des Zeitbegriffs* (Marburg Lectures 1925), *Gesamtausgabe*, Bd. 20, Frankfurt a.M.: Vittorio Klostermann 1979.
- Heidegger, M. (1995). *The fundamental concepts of metaphysics: World, finitude, solitude*. Indiana University Press. Translated by William McNeill and Nicholas Walker. German original: *Die Grundbegriffe der Metaphysik: Welt - Endlichkeit - Einsamkeit* (Freiburg Lectures 1929), *Gesamtausgabe*, Bd. 29/30, Frankfurt a.M.: Vittorio Klostermann 1983.
- Heidegger, M. (1996). *Being and time*. State University of New York Press. Translated from the German *Sein und Zeit* by Joan Stambaugh.
- Heidegger, M. (2002). Phenomenological interpretations in connection with Aristotle. An indication of the hermeneutical situation. In J. Van Buren (Ed.), *Supplements: From the earliest essays to being and time and beyond* (pp. 111–146). State University of New York Press. German original: *Phänomenologische Interpretationen zu Aristoteles (Anzeige der hermeneutischen Situation)* (Marburg and Göttingen 1922), *Gesamtausgabe*, Bd. 62, Frankfurt a.M.: Vittorio Klostermann 2005.
- Hein, M., Dür, W., Eisert, J., Raussendorf, R., Van den Nest, M., & Briegel, H. J. (2006). Entanglement in graph states and its applications. In G. Casati, D. Shepelyansky, P. Zoller, & G. Benenti (Eds.), *Quantum computers, algorithms and chaos*. Proceedings of the international school of physics “Enrico Fermi” (Vol. 162, pp. 115–218). IOS Press. arXiv quant-ph/0602096.
- Hein, M., Eisert, J., & Briegel, H. J. (2004). Multiparty entanglement in graph states. *Physical Review A*, 69, 062311.
- Heisenberg, W. (1927). Über den anschaulichen Inhalt der quantentheoretischen Kinematik und Mechanik. *Zeitschrift für Physik*, 43, 172–198.
- Hempel, C., Maier, C., Romero, J., McClean, J., Monz, T., Shen, H., Jurcevic, P., Lanyon, B. P., Love, P., Babbush, R., Aspuru-Guzik, A., Blatt, R., & Roos, C. F. (2018). Quantum chemistry calculations on a trapped-ion quantum simulator. *Physical Review X*, 8, 031022.
- Hensen, B., Bernien, H., Dréau, A., Reiserer, A., Kalb, N., Blok, M., Ruitenberg, J., Vermeulen, R., Schouten, R., Abellán, C., et al. (2015). Loophole-free Bell inequality violation using electron spins separated by 1.3 kilometres. *Nature*, 526(7575), 682–686.
- Hentschel, K. (1996). Die Entdeckung des Zeeman-Effekts. *Physikalische Blätter*, 52(12), 1232–1235.
- Herbert, N. (1982). FLASH—A superluminal communicator based upon a new kind of quantum measurement. *Foundations of Physics*, 12(12), 1171–1179.
- Herrero-Collantes, M., & Garcia-Escartin, J. C. (2017). Quantum random number generators. *Reviews of Modern Physics*, 89, 015004.

- Hersh, R. (1991). Mathematics has a front and a back. *Synthese*, 88, 127–133.
- Häffner, H., Roos, C., & Blatt, R. (2008). Quantum computing with trapped ions. *Physics Reports*, 469, 155–203.
- High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. <https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai>
- Hilgevoord, J., & Uffink, J. (2024). The uncertainty principle. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2024 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2024/entries/qt-uncertainty/>
- Hájek, A. (2008). Arguments for, or against, probabilism? *British Journal for the Philosophy of Science*, 59(4), 793–819.
- Hofmann, C., Bray, A., Koch, W., Ni, H., & Shvetsov-Shilovski, N. I. (2021). Quantum battles in attoscience: tunnelling. *The European Physical Journal D*, 75(7), 208.
- Hornberger, K., Gerlich, S., Haslinger, P., Nimmrichter, S., & Arndt, M. (2012). Colloquium: Quantum interference of clusters and molecules. *Reviews of Modern Physics*, 84, 157–173.
- Horodecki, R., Horodecki, P., Horodecki, M., & Horodecki, K. (2009). Quantum entanglement. *Reviews of Modern Physics*, 81, 865–942.
- Horton, T. E., Chakraborty, A., & Amant, R. S. (2012). Affordances for robots: a brief survey. *Avant*, 3(2), 70–84.
- Howard, D. (2007). Reduction and emergence in the physical sciences: Some lessons from the particle physics and condensed matter debate. In N. Murphy & W. R. Stoeger (Eds.), *Evolution and emergence: Systems, organisms, persons* (pp. 141–157). Oxford University Press.
- Häse, F., Roch, L. M., & Aspuru-Guzik, A. (2019). Next-generation experimentation with self-driving laboratories. *Trends in Chemistry*, 1, 282–291.
- Humberstone, L. (1992). Direction of fit. *Mind*, 101(401), 59–84.
- Hume, D. (1748). *An enquiry concerning human understanding*. A. Millar. References are to the edition by Tom L. Beauchamp, New York: Oxford University Press, 2000.
- Humphreys, P. (2015). More is different ... sometimes: Ising models, emergence, and undecidability. In B. Falkenburg & M. Morrison (Eds.), *Why more is different* (pp. 137–152). Springer.
- Humphreys, P. (2016). *Emergence: A philosophical account*. Oxford University Press.
- Hurwitz, R. (1992). Milestones in magnetic resonance: Rabi et al. (1938), ‘A new method of measuring nuclear magnetic moment’. *Journal of Magnetic Resonance Imaging*, 2(2), 131–133.
- Ichikawa, J. J., & Steup, M. (2024). The analysis of knowledge. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition. <https://plato.stanford.edu/archives/fall2024/entries/knowledge-analysis/>
- IDQ. (2022). ID Quantique website. <https://www.idquantique.com/>
- Ismael, J. (2025). Quantum mechanics. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2025 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2025/entries/qm/>
- Jaksch, D., Briegel, H.-J., Cirac, J. I., Gardiner, C. W., & Zoller, P. (1999). Entanglement of atoms via cold controlled collisions. *Physical Review Letters*, 82, 1975–1978.
- Jerbi, S., Fiderer, L. J., Poulsen Nautrup, H., Kübler, J. M., Briegel, H. J., & Dunjko, V. (2023). Quantum machine learning beyond kernel methods. *Nature Communications*, 14, 517.
- Jerbi, S., Gyurik, C., Marshall, S., Briegel, H., & Dunjko, V. (2021a). Parametrized quantum policies for reinforcement learning. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P. Liang, & J. Wortman Vaughan (Eds.), *Advances in neural information processing systems* (vol. 34, pp. 28362–28375). Neural Information Processing Systems Foundation.
- Jerbi, S., Trenkwalder, L. M., Poulsen Nautrup, H., Briegel, H. J., & Dunjko, V. (2021b). Quantum enhancements for deep reinforcement learning in large spaces. *Physical Review X Quantum*, 2, 010328.
- Joo, J., Alba, E., García-Ripoll, J. J., & Spiller, T. P. (2013). Generating and verifying graph states for fault-tolerant topological measurement-based quantum computing in two-dimensional optical lattices. *Physical Review A*, 88, 012328.

- Jumper, J., Evans, R., Pritzel, A., Green, T., Figurnov, M., Ronneberger, O., Tunyasuvunakool, K., Bates, R., Židek, A., Potapenko, A., et al. (2021). Highly accurate protein structure prediction with AlphaFold. *Nature*, 596, 583–589.
- Kaelbling, L. P. (1994). Associative reinforcement learning: Functions ink-DNF. *Machine Learning*, 15(3), 279–298.
- Kaelbling, L. P., Littman, M. L., & Moore, A. W. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4(1), 237–285.
- Kane, R. (1998). *The significance of free will*. Oxford University Press.
- Kane, R. (2024). *The complex tapestry of free will: A philosophical odyssey*. Oxford University Press.
- Kant, I. (1781). *Kritik der reinen Vernunft*. Hartknoch. English translation by Paul Guyer and Allen W. Wood: *Critique of Pure Reason*, Cambridge: Cambridge University Press 1998. Following custom, references are to the original pagination of the first edition (A) of 1781 and of the second edition (B) of 1787.
- Keet, A., Fortescue, B., Markham, D., & Sanders, B. C. (2010). Quantum secret sharing with qudit graph states. *Physical Review A*, 82, 062315.
- Keil, G. (2023). Why Frankfurtian all-in can'ts are irrelevant to free will. *Inquiry*, 66(1), 143–164.
- Kempe, J. (2003). Quantum random walks: An introductory overview. *Contemporary Physics*, 44(4), 307–327.
- Kewming, M. J., Shrapnel, S., & Milburn, G. J. (2021). Designing a physical quantum agent. *Physical Review A*, 103, 032411.
- Kim, J. (1998). *Mind in a physical world*. MIT Press.
- Knuuttila, T., & Merz, M. (2009). Understanding by modeling. In H. W. De Regt, S. Leonelli, & K. Eigner (Eds.), *Scientific understanding. Philosophical perspectives*. University of Pittsburgh Press.
- Kochen, S., & Specker, E. (1967). The problem of hidden variables in quantum mechanics. *Journal of Mathematics and Mechanics*, 17, 59–87.
- Kokail, C., Maier, C., van Bijnen, R., Brydges, T., Joshi, M. K., Jurcevic, P., Muschik, C. A., Silvi, P., Blatt, R., Roos, C. F., & Zoller, P. (2019). Self-verifying variational quantum simulation of lattice models. *Nature*, 569(7756), 355–360.
- Krenn, M., Malik, M., Fickler, R., Lapkiewicz, R., & Zeilinger, A. (2016). Automated search for new quantum experiments. *Physical Review Letters*, 116, 090405.
- Krinner, S., Lacroix, N., Remm, A., Di Paolo, A., Genois, E., Leroux, C., Hellings, C., Lazar, S., Swiadek, F., Herrmann, J., Norris, G. J., Andersen, C. K., Müller, M., Blais, A., Eichler, C., & Wallraff, A. (2022). Realizing repeated quantum error correction in a distance-three surface code. *Nature*, 605(7911), 669–674.
- Kuhn, T. S. (1962). *The structure of scientific revolutions* (4th ed.). University of Chicago Press. With an introduction by Ian Hacking, 2012.
- Ladyman, J. (2023). Structural realism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2023 edition. <https://plato.stanford.edu/archives/sum2023/entries/structural-realism/>
- Ladyman, J., & Ross, D. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press. Co-authored with David Spurrett and John Collie.
- Landsman, K. (2017). *Foundations of quantum theory. From classical concepts to operator algebras*. Springer.
- Langford, J., & Zhang, T. (2007). The epoch-greedy algorithm for multi-armed bandits with side information. In J. Platt, D. Koller, Y. Singer, & S. Roweis (Eds.), *Advances in neural information processing systems* (Vol. 20). Neural Information Processing Systems Foundation.
- Laplace, P. S. (1820). *Théorie Analytique des Probabilités*. V. Courcier. English translation by F. W. Truscott and F. L. Emory: *A philosophical essay on probabilities*, New York, NY: Dover, 1951. Page references are to this translation.
- Laudan, L. (1981). A confutation of convergent realism. *Philosophy of Science*, 48(1), 19–49.
- Lawson, R. (2006). The science of cycology: Failures to understand how everyday objects work. *Memory & Cognition*, 34, 1667–1675.

- Leach, J., Padgett, M. J., Barnett, S. M., Franke-Arnold, S., & Courtial, J. (2002). Measuring the orbital angular momentum of a single photon. *Physical Review Letters*, 88, 257901.
- Lear, J. (1984). The disappearing “we”. *Proceedings of the Aristotelian Society, Supplementary Volumes*, 58, 219–258.
- LeBlanc, F. (2010). *An introduction to stellar astrophysics*. Wiley.
- LeMaitre, P. A., Krumm, M., & Briegel, H. J. (2024). Multi-excitation projective simulation with a many-body physics inspired inductive bias. arXiv 2402.10192.
- Levine, M. P. (2011). Philosophers on miracles. In G. H. Twelftree (Ed.), *The Cambridge companion to miracles* (pp. 291–308). Cambridge University Press.
- Lewes, G. H. (1875). *Problems of life and mind* (vol. 2). Kegan Paul, Trench, Turbner.
- Lewis, D. K. (1986a). *On the plurality of worlds*. Blackwell.
- Lewis, D. K. (1986b). *Philosophical Papers* (vol. 2). Oxford University Press.
- Lewis, D. K. (1996). Elusive knowledge. *Australasian Journal of Philosophy*, 74(4), 549–567.
- Lewis, R. L., Howes, A., & Singh, S. (2014). Computational rationality: Linking mechanism and behavior through bounded utility maximization. *Topics in Cognitive Science*, 6(2), 279–311.
- Libbrecht, K. G. (2021). *Snow crystals: A case study in sponaneous structure formation*. Princeton University Press.
- Lidar, D. A., & Brun, T. A. (Eds.). (2013). *Quantum error correction*. Cambridge University Press.
- Lindblad, G. (1976). On the generators of quantum dynamical semigroups. *Communications in Mathematical Physics*, 48(2), 119–130.
- Linden, N., & Popescu, S. (1998). On multi-particle entanglement. *Fortschritte der Physik*, 46(4–5), 567–578.
- Linson, A., Clark, A., Ramamoorthy, S., & Friston, K. (2018). The active inference approach to ecological perception: General information dynamics for natural and artificial embodied cognition. *Frontiers in Robotics and AI*, 5, 21.
- Lockwood, M. (1996). ‘Many minds’ interpretations of quantum mechanics. *The British Journal for the Philosophy of Science*, 47(2), 159–188.
- López-Incera, A., Nouvian, M., Ried, K., Müller, T., & Briegel, H. J. (2021). Honeybee communication during collective defence is shaped by predation. *BMC Biology*, 19(1), 106.
- López-Incera, A., Ried, K., Müller, T., & Briegel, H. J. (2020). Development of swarm behavior in artificial learning agents that adapt to different foraging environments. *PLoS One*, 15(12), e0243628.
- Löwe, B., & Müller, T. (2011). Data and phenomena in conceptual modelling. *Synthese*, 182(1), 131–148.
- Lukens, J. M., Law, K. J. H., Jasra, A., & Lougovski, P. (2020). A practical and efficient approach for Bayesian quantum state estimation. *New Journal of Physics*, 22(6), 063038.
- Ma, X.-S., Kofler, J., & Zeilinger, A. (2016). Delayed-choice gedanken experiments and their realizations. *Reviews of Modern Physics*, 88, 015005.
- Ma-Kellams, C. (2015). When perceiving the supernatural changes the natural: Religion and agency detection. *Journal of Cognition and Culture*, 15(3–4), 337–343.
- Mackenzie, D. (2009). The fifty-one percent solution. In D. Mackenzie (Ed.), *What’s happening in the mathematical sciences?* (vol. 7, pp. 34–45). American Mathematical Society.
- Mackie, J. L. (1977). *Ethics: Inventing right and wrong*. Penguin.
- Magniez, F., Nayak, A., Roland, J., & Santha, M. (2011). Search via quantum walk. *SIAM Journal on Computing*, 40(1), 142–164.
- Mair, A., Vaziri, A., Weihs, G., & Zeilinger, A. (2001). Entanglement of the orbital angular momentum states of photons. *Nature*, 412(6844), 313–316.
- Majumder, A., Krumm, M., Radkohl, T., Poulsen Nautrup, H., Jerbi, S., & Briegel, H. J. (2024). Variational measurement-based quantum computation for generative modeling. *Physical Review A*, 110, 062616.
- Malik, M., & Boyd, R. W. (2014). Quantum imaging technologies. *La Rivista del Nuovo Cimento*, 37(5), 273–332.
- Malik, M., Erhard, M., Huber, M., Krenn, M., Fickler, R., & Zeilinger, A. (2016). Multi-photon entanglement in high dimensions. *Nature Photonics*, 10(4), 248–252.

- Mannalath, V., Mishra, S., & Pathak, A. (2023). A comprehensive review of quantum random number generators: Concepts, classification and the origin of randomness. *Quantum Information Processing*, 22, 439.
- Margalit, Y., Dobkowski, O., Zhou, Z., Amit, O., Japha, Y., Moukouri, S., Rohrlach, D., Mazumdar, A., Bose, S., Henkel, C., & Folman, R. (2021). Realization of a complete Stern-Gerlach interferometer: Toward a test of quantum gravity. *Science Advances*, 7(22), eabg2879.
- Marov, M. Y. (2015). The structure of the universe. In M. Y. Marov (Ed.), *The fundamentals of modern astrophysics: A survey of the cosmos from the home planet to space frontiers* (pp. 279–294). Springer.
- Maudlin, T. (2019). *Philosophy of physics. Quantum theory*. Princeton University Press.
- Mautner, J., Makmal, A., Manzano, D., Tiersch, M., & Briegel, H. J. (2015). Projective simulation for classical learning agents: A comprehensive investigation. *New Generation Computing*, 33, 69–114.
- McClean, J. R., Romero, J., Babbush, R., & Aspuru-Guzik, A. (2016). The theory of variational hybrid quantum-classical algorithms. *New Journal of Physics*, 18(2), 023023.
- McDowell, J. (2013). The myth of the mind as detached. In J. K. Schear (Ed.), *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.
- McDowell, J. H. (1994). *Mind and world*. Harvard University Press.
- McGrew, T., & Larmer, R. (2024). Miracles. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2024 edition. <https://plato.stanford.edu/archives/sum2024/entries/miracles/>
- Meincke, A. S. (2024). Emergent properties. In A. R. J. Fisher & A.-S. Maurin (Eds.), *Routledge handbook of properties* (pp. 347–357). Routledge.
- Melnikov, A. A., Makmal, A., & Briegel, H. J. (2018a). Benchmarking projective simulation in navigation problems. *IEEE Access*, 6, 64639–64648.
- Melnikov, A. A., Poulsen Nautrup, H., Krenn, M., Dunjko, V., Tiersch, M., Zeilinger, A., & Briegel, H. J. (2018b). Active learning machine learns to create new quantum experiments. *Proceedings of the National Academy of Sciences*, 115, 1221–1226.
- Melnikov, A. A., Makmal, A., Dunjko, V., & Briegel, H. J. (2017). Projective simulation with generalization. *Scientific Reports*, 7(14430).
- Merleau-Ponty, M. (1945). *Phénoménologie de la perception*. Éditions Gallimard. English translation by C. Smith: *Phenomenology of Perception*. Routledge, London 1962. Page references are to this translation.
- Mermin, N. D. (1981). Quantum mysteries for anyone. *Journal of Philosophy*, 78(7), 397–408.
- Mermin, N. D. (1990). Quantum mysteries revisited. *American Journal of Physics*, 58, 731–734.
- Mermin, N. D. (2017). Why QBism is not the Copenhagen interpretation and what John Bell might have thought of it. In R. Bertlmann & A. Zeilinger (Eds.), *Quantum [Un]speakables II: Half a century of bell's theorem* (pp. 83–93). Springer.
- Michelson, A. A. (1903). *Light waves and their uses*. University of Chicago Press.
- Milgram, E. (2005). *Ethics done right*. Cambridge University Press.
- Miller, J., Sanders, S., & Miyake, A. (2017). Quantum supremacy in constant-time measurement-based computation: A unified architecture for sampling and verification. *Physical Review A*, 96, 062320.
- Minsky, M. (1961). Steps toward artificial intelligence. *Proceedings of the IRE*, 49(1), 8–30.
- Minsky, M. (1967). *Computation: Finite and infinite machines*. Prentice-Hall.
- Mitchell, K. J. (2023). *Free agents. How evolution gave us free will*. Princeton University Press.
- Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., Graves, A., Riedmiller, M., Fidjeland, A. K., Ostrovski, G., Petersen, S., Beattie, C., Sadik, A., Antonoglou, I., King, H., Kumaran, D., Wierstra, D., Legg, S., & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518(7540), 529–533.
- Molina-Terriza, G., Torres, J. P., & Torner, L. (2007). Twisted photons. *Nature Physics*, 3(5), 305–310.
- Montague, M. (2007). Against propositionalism. *Noûs*, 41(3), 503–518.
- Morales, J. (2023). Emergent agent causation. *Synthese*, 201, 138.

- Moretti, L., & Piazza, T. (2018). Defeaters in current epistemology: Introduction to the special issue. *Synthese*, 195, 2845–2854.
- Morgan, M. S., & Morrison, M. (1999). *Models as mediators*. Cambridge University Press.
- Mulder, J., Müller, T., Ometto, D., & Van Miltenburg, N. (2022). Causality and determination, powers and agency: Anscombean perspectives. *Synthese*, 200:452. Introduction to a topical collection consisting of 12 articles.
- Mulder, J. M. (2016). A vital challenge to materialism. *Philosophy*, 91, 153–182.
- Mulder, J. M. (2021). Animals run about the world in all sorts of paths: Varieties of indeterminism. *Synthese*, 199, 11945–11961.
- Müller, T. (2022). “Let’s build an Anscombe box”: assessing Anscombe’s rebuttal of the statistics objection against indeterminism-based free agency. *Synthese*, 200, 74.
- Müller, T. (2023). Taming Pereboom’s wild coincidences. *Mind*, 132(527), 789–802.
- Müller, T., & Briegel, H. J. (2018). A stochastic process model for free agency under indeterminism. *Dialectica*, 72(2), 219–252.
- Müller, T., & Placek, T. (2018). Defining determinism. *British Journal for the Philosophy of Science*, 69(1), 215–252.
- Müller, T., & Placek, T. (2023). On the condition of setting independence. *European Journal for the Philosophy of Science*, 13, 53.
- Myrvold, W. (2022). Philosophical Issues in Quantum Theory. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Fall 2022 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/fall2022/entries/qt-issues/>.
- Myrvold, W., Genovese, M., & Shimony, A. (2024). Bell’s Theorem. In E. N. Zalta, & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy* (Spring 2024 ed.). Metaphysics Research Lab, Stanford University. <https://plato.stanford.edu/archives/spr2024/entries/bell-theorem/>.
- Nagel, T. (1986). *The view from nowhere*. Oxford University Press.
- Neuman, K. C., & Nagy, A. (2008). Single-molecule force spectroscopy: optical tweezers, magnetic tweezers and atomic force microscopy. *Nature Methods*, 5(6), 491–505.
- Nguyen, P. D. H., Georgie, Y. K., Kayhan, E., Eppl, M., Hafner, V. V., & Wermter, S. (2021). Sensorimotor representation learning for an “active self” in robots: A model survey. *Künstliche Intelligenz*, 32, 9–35.
- Nicholson, D. J., & Dupré, J. (Eds.) (2018). *Everything flows: Towards a processual philosophy of biology*. Oxford University Press.
- Nielsen, M. A., & Chuang, I. L. (2010). *Quantum computation and quantum information* (10th anniversary ed.). Cambridge University Press.
- Nigg, D., Müller, M., Martinez, E. A., Schindler, P., Hennrich, M., Monz, T., Martin-Delgado, M. A., & Blatt, R. (2014). Quantum computations on a topologically encoded qubit. *Science*, 345(6194), 302–305.
- NIST. (2018). NIST’s new quantum method generates really random numbers. <https://www.nist.gov/news-events/news/2018/04/nists-new-quantum-method-generates-really-random-numbers>.
- NIST. (2022). <https://csrc.nist.gov/projects/interoperable-randomness-beacons>.
- Noble, R., & Noble, D. (2021). Can reasons and values influence action: How might intentional agency work physiologically? *Journal for General Philosophy of Science*, 52(2), 277–295.
- Norman, D. A. (2013). *The design of everyday things*. Basic Books. Revised and expanded edition.
- Norton, J. D. (2008). The Dome: An unexpectedly simple failure of determinism. *Philosophy of Science*, 75(5), 786–798.
- Obenchain, T. G. (2016). *Genius belabored: Childbed fever and the tragic life of Ignaz Semmelweis*. The University of Alabama Press.
- O’Connor, T. (2021). Emergent properties. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Winter 2021 edition. <https://plato.stanford.edu/archives/win2021/entries/properties-emergent/>.
- Odegard, V. H., & Schatz, D. G. (2006). Targeting of somatic hypermutation. *Nature Reviews Immunology*, 6(8), 573–583.

- Øhrstrøm, P., & Hasle, P. (2024). Future contingents. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2024 edition. <https://plato.stanford.edu/archives/fall2024/entries/future-contingents/>
- Øhrstrøm, P., & Jakobsen, D. (2022). Ockhamistic inspiration in modern tense-logic. In A. Santelli (Ed.), *Ockhamism and philosophy of time: Semantic and metaphysical issues concerning future contingents* (pp. 33–54). Springer.
- Okasha, S. (2018). *Agents and goals in evolution*. Oxford University Press.
- Okasha, S. (2024). The concept of agent in biology: Motivations and meanings. *Biological Theory*, 19, 6–10.
- O'Malley, P. J. J., Babbush, R., Kivlichan, I. D., Romero, J., McClean, J. R., Barends, R., Kelly, J., Roushan, P., Tranter, A., Ding, N., Campbell, B., Chen, Y., Chen, Z., Chiaro, B., Dunsworth, A., Fowler, A. G., Jeffrey, E., Lucero, E., Megrant, A., Mutus, J. Y., Neeley, M., Neill, C., Quintana, C., Sank, D., Vainsencher, A., Wenner, J., White, T. C., Coveney, P. V., Love, P. J., Neven, H., Aspuru-Guzik, A., & Martinis, J. M. (2016). Scalable quantum simulation of molecular energies. *Physical Review X*, 6, 031007.
- Oreskes, N. (2019). *Why trust science?* Princeton University Press.
- Orsucci, D., Briegel, H. J., & Dunjko, V. (2018). Faster quantum mixing for slowly evolving sequences of Markov chains. *Quantum*, 2, 105.
- Oudeyer, P.-Y., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurobotics*, 1, 6.
- Paparo, G. D., Dunjko, V., Makmal, A., Martin-Delgado, M. A., & Briegel, H. J. (2014). Quantum speedup for active learning agents. *Physical Review X*, 4(3), 031002.
- Paraniak, M. M., & Englert, B.-G. (2021). Quantum dynamical simulation of a transversal Stern-Gerlach interferometer. *Symmetry*, 13(9), 1660.
- Paris, M. G. A., & Řeháček, J. (Eds.). (2004). *Quantum state estimation. Lecture notes in physics* (vol. 649). Springer.
- Pauli, W. (1925). Über den Zusammenhang des Abschlusses der Elektronengruppen im Atom mit der Komplexstruktur der Spektren. *Zeitschrift für Physik*, 31, 765–783.
- Pauli, W. (1927). Zur Quantenmechanik des magnetischen Elektrons. *Zeitschrift für Physik*, 43, 601–623.
- Pavičić, M., Merlet, J.-P., McKay, B., & Megill, N. D. (2005). Kochen-Specker vectors. *Journal of Physics A: Mathematical and General*, 38(7), 1577.
- Pazem, J., Krumm, M., Vining, A. Q., Fiderer, L. J., & Briegel, H. J. (2024). Free energy projective simulation (feps), Active inference with interpretability. arXiv 2411.14991.
- Pearl, J. (2002). Reasoning with cause and effect. *AI Magazine*, 23(1), 95–111.
- Pearl, J., & Mackenzie, D. (2018). *The book of why*. Basic Books.
- Pechukas, P. (1994). Reduced dynamics need not be completely positive. *Physical Review Letters*, 73, 1060–1062.
- Pereboom, D. (2001). *Living without free will*. Cambridge University Press.
- Pereboom, D. (2014). *Free will, agency, and meaning in life*. Oxford University Press.
- Peres, A. (1978). Unperformed experiments have no results. *American Journal of Physics*, 46, 745–747.
- Peres, A. (1991). Two simple proofs of the Kochen-Specker theorem. *Journal of Physics A: Mathematical and General*, 24(4), L175.
- Peres, A. (2002). *Quantum theory: Concepts and methods*. Fundamental theories of physics (Vol. 57). Kluwer Academic Publishers.
- Pesic, P. (1999). Wrestling with Proteus: Francis Bacon and the “torture” of nature. *Isis*, 90(1), 84–94.
- Peters, J., Janzing, D., & Schölkopf, B. (2017). *Elements of causal inference. Foundations and learning algorithms*. MIT Press.
- Pettigrew, R. (2016). *Accuracy and the laws of credence*. Oxford University Press.
- Phipps, T. E., & Stern, O. (1932). Über die Einstellung der Richtungsquantelung. *Zeitschrift für Physik*, 73, 185–191.
- Pickard, H. (2021). Addiction and the self. *Noûs*, 55, 737–761.

- Pienaar, J. (2020). Extending the agent in QBism. *Foundations of Physics*, 50, 1894–1920.
- Pienaar, J. (2021). QBism and relational quantum mechanics compared. *Foundations of Physics*, 51(5), 1–18.
- Pironio, S., Acín, A., Massar, S., de la Giroday, A. B., Matsukevich, D. N. and 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.
- Pombo, O. (2011). Neurath and the encyclopaedic project of unity of science. In J. Symons, O. Pombo, & J. M. Torres (Eds.), *Otto Neurath and the unity of science* (pp. 59–70). Springer.
- Popescu, S. (1994). Bell's inequalities versus teleportation: What is nonlocality? *Physical Review Letters*, 72, 797–799.
- Poulsen Nautrup, H., Delfosse, N., Dunjko, V., Briegel, H. J., & Friis, N. (2019). Optimizing quantum error correction codes with reinforcement learning. *Quantum*, 3, 215.
- Preskill, J. (2018). Quantum computing in the NISQ era and beyond. *Quantum*, 2, 79.
- Putnam, H. (1967). Time and physical geometry. *The Journal of Philosophy*, 64(8), 240–247.
- Quine, W. V. O. (1948). On what there is. *Review of Metaphysics*, 2(5), 21–38.
- Rabi, I. I., Millman, S., Kusch, P., & Zacharias, J. R. (1939). The molecular beam resonance method for measuring nuclear magnetic moments. The magnetic moments of ${}^3\text{Li}^6$, ${}^3\text{Li}^7$ and ${}^9\text{F}^{19}$. *Physical Review*, 55, 526–535.
- Rabi, I. I., Zacharias, J. R., Millman, S., & Kusch, P. (1938). A new method of measuring nuclear magnetic moment. *Physical Review*, 53, 318.
- Ragab, H., Milburn, A., Razavi, K., Bos, H., & Giuffrida, C. (2021). Crosstalk: Speculative data leaks across cores are real. In *2021 IEEE symposium on security and privacy (SP)* (pp. 1852–1867).
- Rapp, J. T., Bremer, B. J., & Romero, P. A. (2024). Self-driving laboratories to autonomously navigate the protein fitness landscape. *Nature Chemical Engineering*, 1(1), 97–107.
- Rauch, H. (1998). Quantum phenomena studied by neutron interferometry. *Physica Scripta*, 1998(T76), 24.
- Rauch, H., Zeilinger, A., Badurek, G., Wilfing, A., Bauspiess, W., & Bonse, U. (1975). Verification of coherent spinor rotation of fermions. *Physics Letters A*, 54(6), 425–427.
- Raussendorf, R., & Briegel, H. J. (2001). A one-way quantum computer. *Physical Review Letters*, 86, 5188–5191.
- Raussendorf, R., Browne, D. E., & Briegel, H. J. (2003). Measurement-based quantum computation on cluster states. *Physical Review A*, 68, 022312.
- Reck, M., Zeilinger, A., Bernstein, H. J., & Bertani, P. (1994). Experimental realization of any discrete unitary operator. *Physical Review Letters*, 73, 58–61.
- Renes, J. M., Blume-Kohout, R., Scott, A. J., & Caves, C. M. (2004). Symmetric informationally complete quantum measurements. *Journal of Mathematical Physics*, 45, 2171–2180.
- Rheinberger, H.-J. (2021a). On the narrative order of experimentation. In M. Carrier, R. Mertens, & C. Reinhardt (Eds.), *Narratives and comparisons: Adversaries or allies in understanding science?* Bielefeld University Press.
- Rheinberger, H.-J. (2021b). *Spalt und Fuge. Eine Phänomenologie des Experiments*. Suhrkamp.
- Ried, K., Müller, T., & Briegel, H. J. (2019). Modelling collective motion based on the principle of agency: General framework and the case of marching locusts. *PLoS ONE*, 14(2), 1–21.
- Rietdijk, C. (1966). A rigorous proof of determinism derived from the special theory of relativity. *Philosophy of Science*, 33(4), 341–344.
- Robb, D., Heil, J., & Gibb, S. (2023). Mental causation. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford Encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Spring 2023 edition. <https://plato.stanford.edu/archives/spr2023/entries/mental-causation/>
- Rumberg, A. (2016). *Transitions toward a semantics for real possibility*. Ph.D. Thesis, University of Utrecht.
- Russell, B. (1912). On the notion of cause. *Proceedings of the Aristotelian Society*, 13, 1–26.
- Russell, B. (1948). *Human knowledge: Its scope and limits*. Allen and Unwin.
- Russell, S., & Norvig, P. (2022). *Artificial intelligence: A modern approach* (4th ed.). Pearson Education.

- Russo, F. (2022). *Techno-scientific practices. An informational approach*. Rowman & Littlefield.
- Ryle, G. (1949). *The concept of mind*. Hutchinson.
- Saffman, M. (2016). Quantum computing with atomic qubits and Rydberg interactions: progress and challenges. *Journal of Physics B: Atomic, Molecular and Optical Physics*, 49(20), 202001.
- Saggio, V., Asenbeck, B. E., Hamann, A., Strömberg, T., Schiansky, P., Dunjko, V., Friis, N., Harris, N. C., Hochberg, M., Englund, D., Wölk, S., Briegel, H. J., & Walther, P. (2021). Experimental quantum speed-up in reinforcement learning agents. *Nature*, 591(7849), 229–233.
- Sakurai, J. J., & Napolitano, J. (2020). *Modern quantum mechanics* (3rd ed.). Cambridge University Press.
- Salge, C., Glackin, C., & Polani, D. (2014). Empowerment – An introduction. In M. Prokopenko (Ed.), *Guided self-organization: Inception* (pp. 67–114). Springer.
- Salge, C., & Guckelsberger, C. (2016). Does empowerment maximisation allow for enactive artificial agents? In *Proceedings of the Artificial Life Conference 2016* (pp. 704–711). MIT Press.
- Sartenaer, O., & Baysan, U. (2021). Non-standard approaches to emergence: Introduction to the special issue. *Synthese*, 199, 7773–7776.
- Sartre, J.-P. (1943). *L'Être et le néant*. Gallimard. English translation by Sarah Richmond: *Being and Nothingness*, Routledge, London 2018.
- Scheur, J. K., editor (2013). *Mind, reason, and being-in-the-world: The McDowell-Dreyfus debate*. Routledge.
- Schindler, P., Nigg, D., Monz, T., Barreiro, J. T., Martinez, E., Wang, S. X., Quint, S., Brandl, M. F., Nebendahl, V., Roos, C. F., Chwalla, M., Hennrich, M., & Blatt, R. (2013). A quantum information processor with trapped ions. *New Journal of Physics*, 15(12), 123012.
- Schlingemann, D. (2002). Stabilizer codes can be realized as graph codes. *Quantum Information and Computation*, 2(4), 307–323.
- Schmidt-Böcking, H., Gruber, G., & Friedrich, B. (2023). One hundred years ago Alfred Landé unriddled the anomalous Zeeman effect and presaged electron spin. *Physica Scripta*, 98(1), 014005.
- Schmidt-Böcking, H., Schmidt, L., Lüdde, H. J., Trageser, W., Templeton, A., & Sauer, T. (2016). The Stern-Gerlach experiment revisited. *The European Physical Journal H*, 41, 327–364.
- Schrödinger, E. (1935). Die gegenwärtige Situation in der Quantenmechanik. *Naturwissenschaften*, 23, 807–812, 823–828, 844–849. English translation by J. D. Trimmer, “The Present Situation in Quantum Mechanics”, *Proceedings of the American Philosophical Society* 124(5), 1980, pp. 323–338.
- Schuld, M., Sinayskiy, I., & Petruccione, F. (2015). An introduction to quantum machine learning. *Contemporary Physics*, 56(2), 172–185.
- Schwinger, J. (2001). In B.-G. Englert (Ed.), *Quantum mechanics. Symbolism of atomic measurements*. Springer.
- Schwinger, J., Scully, M. O., & Englert, B.-G. (1988). Is spin coherence like Humpty-Dumpty? II. General theory. *Zeitschrift für Physik D. Atoms, Molecules and Clusters*, 10, 135–144.
- Scully, M. O., & Drühl, K. (1982). Quantum eraser: A proposed photon correlation experiment concerning observation and “delayed choice” in quantum mechanics. *Physical Review A*, 25, 2208–2213.
- Scully, M. O., Englert, B.-G., & Schwinger, J. (1989). Spin coherence and Humpty-Dumpty. III. The effects of observation. *Physical Review A*, 40, 1775–1784.
- Scully, M. O., Englert, B.-G., & Walther, H. (1991). Quantum optical tests of complementarity. *Nature*, 351(6322), 111–116.
- Searle, J. (1983). *Intentionality: An essay in the philosophy of mind*. Cambridge University Press.
- Senior, A. W., Evans, R., Jumper, J., Kirkpatrick, J., Sifre, L., Green, T., Qin, C., Žídek, A., Nelson, A. W. R., Bridgland, A., Penadones, H., Petersen, S., Simonyan, K., Crossan, S., Kohli, P., Jones, D. T., Silver, D., Kavukcuoglu, K., & Hassabis, D. (2020). Improved protein structure prediction using potentials from deep learning. *Nature*, 577(7792), 706–710.
- Shalm, L. K., Meyer-Scott, E., Christensen, B. G., Bierhorst, P., Wayne, M. A., Stevens, M. J., Gerrits, T., Glancy, S., Hamel, D. R., Allman, M. S., Coakley, K. J., Dyer, S. D., Hodge, C.,

- Lita, A. E., Verma, V. B., Lambrocco, C., Tortorici, E., Migdall, A. L., Zhang, Y., Kumor, D. R., Farr, W. H., Marsili, F., Shaw, M. D., Stern, J. A., Abellán, C., Amaya, W., Pruneri, V., Jennewein, T., Mitchell, M. W., Kwiatt, P. G., Bienfang, J. C., Mirin, R. P., Knill, E., & Nam, S. W. (2015). Strong loophole-free test of local realism. *Physical Review Letters*, 115, 250402.
- Silver, D., Hubert, T., Schrittwieser, J., Antonoglou, I., Lai, M., Guez, A., Lanctot, M., Sifre, L., Kumaran, D., Graepel, T., Lillicrap, T., Simonyan, K., & Hassabis, D. (2018). A general reinforcement learning algorithm that masters chess, shogi, and Go through self-play. *Science*, 362(6419), 1140–1144.
- Skotiniotis, M., Dür, W., & Sekatski, P. (2017). Macroscopic superpositions require tremendous measurement devices. *Quantum*, 1, 34.
- Smith, I. D., Krumm, M., Fiderer, L. J., Poulsen Nautrup, H., & Briegel, H. J. (2023). The min-entropy of classical-quantum combs for measurement-based applications. *Quantum*, 7, 1206.
- Stacey, B. C. (2019). Ideas abandoned en route to QBism. arXiv 1911.07386.
- Steward, H. (2012). *A metaphysics for freedom*. Oxford University Press.
- Steward, H. (2020). Free will and external reality: Two scepticisms compared. *Proceedings of the Aristotelian Society*, 120(1), 1–20.
- Steward, H. (2021). What is determinism? Why we should ditch the entailment definition. In M. Hausmann & J. Noller (Eds.), *Free will: Historical and analytic perspectives* (pp. 17–43). Springer.
- Steward, H. (2022). Animal agency. In L. Ferrero (Ed.), *The Routledge handbook of philosophy of agency* (pp. 101–108). Routledge.
- Straub, I. (2016). ‘It looks like a human!’ The interrelation of social presence, interaction and agency ascription: A case study about the effects of an android robot on social agency ascription. *AI & Society*, 31(4), 553–571.
- Strawson, P. F. (1959). *Individuals. An essay in descriptive metaphysics*. Routledge.
- Strawson, P. F. (1962). Freedom and resentment. *Proceedings of the British Academy*, 48, 187–211.
- Strawson, P. F. (1992). *Analysis and metaphysics*. Oxford University Press.
- Stricker, R., Meth, M., Postler, L., Edmunds, C., Ferrie, C., Blatt, R., Schindler, P., Monz, T., Kueng, R., & Ringbauer, M. (2022). Experimental single-setting quantum state tomography. *Physical Review X Quantum*, 3, 040310.
- Sutton, R. S., & Barto, A. G. (2018). *Reinforcement learning: An introduction* (2nd ed.). MIT Press.
- Szegedy, M. (2004). Quantum speed-up of Markov chain based algorithms. In *45th annual IEEE symposium on foundations of computer science* (pp. 32–41).
- Tennant, N. (2023). Logicism and neologicism. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Winter 2023 edition. <https://plato.stanford.edu/archives/win2023/entries/logicism/>
- Terhal, B. M. (2015). Quantum error correction for quantum memories. *Reviews of Modern Physics*, 87, 307–346.
- Thomas, L. (1926). The motion of the spinning electron. *Nature*, 117(2945), 514.
- Thomason, R. H., & Horty, J. (2022). Artificial and machine agency. In L. Ferrero (Ed.), *The Routledge handbook of philosophy of agency* (pp. 366–375). Routledge.
- Thompson, M. (2004). What is it to wrong someone? A puzzle about justice. In R. J. Wallace, P. Pettit, S. Scheffler, & M. Smith (Eds.), *Reason and value: Themes from the moral philosophy of Joseph Raz* (pp. 333–384). Oxford University Press.
- Thompson, M. (2008). *Life and action*. Harvard University Press.
- t’Hooft, G. (2016). *The cellular automaton interpretation of quantum mechanics*. Springer.
- Thorne, K. S. (2017). Nobel lecture delivered december 8, 2017: LIGO and gravitational waves III. Nobel Prize Outreach. www.nobelprize.org/prizes/physics/2017/thorne/lecture/.
- Tiersch, M., Ganahl, E., & Briegel, H. J. (2015). Adaptive quantum computation in changing environments using projective simulation. *Scientific Reports*, 5, 12874.
- Timpson, C. G. (2008). Quantum Bayesianism: A study. *Studies in History and Philosophy of Modern Physics*, 39(3), 579–609.

- Tomasello, M. (2022). *The evolution of agency: Behavioral organization from lizards to humans*. MIT Press.
- Trenkwalder, L. M., López-Incera, A., Poulsen Nautrup, H., Flamini, F., & Briegel, H. J. (2023). Automated gadget discovery in the quantum domain. *Machine Learning: Science and Technology*, 4(3), 035043.
- Tulving, E. (1972). Episodic and semantic memory. In E. Tulving, & W. Donaldson (Eds.), *Organization of Memory* (pp. 2799–2804). Academic Press.
- Tumulka, R. (2021). Bohmian mechanics. In E. Knox & A. Wilson (Eds.), *Routledge companion to the philosophy of physics* (chap. 17). Routledge.
- Turing, A. M. (1936). On computable numbers, with an application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society, Series 2*, 42, 230–265.
- Uebel, T. E. (2007). *Empiricism at the crossroads: The Vienna Circle's protocol-sentence debate*. Open Court.
- Uhlenbeck, G. E., & Goudsmit, S. (1925). Ersetzung der Hypothese vom unmechanischen Zwang durch eine Forderung bezüglich des inneren Verhaltens jedes einzelnen Elektrons. *Naturwissenschaften*, 47, 953–954.
- Uhlenbeck, G. E., & Goudsmit, S. (1926). Spinning electrons and the structure of spectra. *Nature*, 117(2938), 264–265.
- Vaidman, L. (2021). Many-worlds interpretation of quantum mechanics. In E. N. Zalta (Ed.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2021 edition. <https://plato.stanford.edu/archives/fall2021/entries/qm-manyworlds/>
- Van Delft, D., & Kes, P. (2010). The discovery of superconductivity. *Physics Today*, 63(9), 38–43.
- Van Helden, A. (1976). The importance of the transit of mercury of 1631. *Journal for the History of Astronomy*, 7(1), 1–10.
- Van Inwagen, P. (1983). *An essay on free will*. Oxford University Press.
- Van Lier, M. (2025). *Modeling the artificial agent*. Ph.D. Thesis, University of Konstanz.
- Van Otterlo, M., & Wiering, M. (2012). Reinforcement learning and Markov decision processes. In M. Van Otterlo, & M. Wiering (Eds.), *Reinforcement learning: State of the art* (pp. 25–52). Springer.
- Van Riel, R., & Van Gulick, R. (2025). Scientific reduction. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Summer 2025 edition. <https://plato.stanford.edu/archives/sum2025/entries/scientific-reduction/>
- Vandersypen, L. M. K., & Chuang, I. L. (2005). NMR techniques for quantum control and computation. *Reviews of Modern Physics*, 76, 1037–1069.
- Velleman, J. D. (1992). What happens when someone acts? *Mind*, 101(403), 461–481.
- Vidal, G. (2000). Entanglement monotones. *Journal of Modern Optics*, 47(2–3), 355–376.
- Vineberg, S. (2022). Dutch book arguments. In E. N. Zalta & U. Nodelman (Eds.), *The Stanford encyclopedia of philosophy*. Metaphysics Research Lab, Stanford University, Fall 2022 edition. <https://plato.stanford.edu/archives/fall2022/entries/dutch-book/>
- Vladimirov, N., & Sourjik, V. (2009). Chemotaxis: How bacteria use memory. *Biological Chemistry*, 390(11), 1097–1104.
- Von Neumann, J. (1932). *Mathematische grundlagen der quantenmechanik*. Julius Springer. English translation by R. T. Beyer: *Mathematical Foundations of Quantum Mechanics*. Princeton University Press, Princeton, NJ 1955.
- Wallnöfer, J., Melnikov, A. A., Dür, W., & Briegel, H. J. (2020). Machine learning for long-distance quantum communication. *Physical Review X Quantum*, 1, 010301.
- Wang, J., Sciarrino, F., Laing, A., & Thompson, M. G. (2020). Integrated photonic quantum technologies. *Nature Photonics*, 14(5), 273–284.
- Watkins, C. J. C. H., & Dayan, P. (1992). Q-learning. *Machine Learning*, 8(3), 279–292.
- Watson, G. (1987). Responsibility and the limits of evil: Variations on a Strawsonian theme. In J. M. Fischer & M. Ravizza (Eds.), *Perspectives on moral responsibility* (pp. 119–148). Cornell University Press.
- Wegner, D. (2002). *The illusion of conscious will*. MIT Press.

- Weimer, H., Müller, M., Lesanovsky, I., Zoller, P., & Büchler, H. P. (2010). A Rydberg quantum simulator. *Nature Physics*, 6(5), 382–388.
- Wellmann, J. (2024). *Biological motion*. Zone Books.
- Werndl, C. (2011). On the observational equivalence of continuous-time deterministic and indeterministic descriptions. *European Journal for Philosophy of Science*, 1, 193–225.
- Werner, R. F. (1989). Quantum states with Einstein-Podolsky-Rosen correlations admitting a hidden-variable model. *Physical Review A*, 40, 4277–4281.
- Werner, S. A., Colella, R., Overhauser, A. W., & Eagen, C. F. (1975). Observation of the phase shift of a neutron due to precession in a magnetic field. *Physical Review Letters*, 35, 1053–1055.
- Wheeler, J. A., & Zurek, W. H. (Eds.). (1984). *Quantum theory and measurement*. Princeton University Press.
- Wheeler, M. (2005). *Reconstructing the cognitive world*. MIT Press.
- Winograd, T., & Flores, F. (1987). *Understanding computers and cognition: A new foundation for design*. Addison Wesley.
- Wittgenstein, L. (1969). *On certainty*. Basil Blackwell. Manuscript dated 1951, translated and edited by Denis Paul and G. E. M. Anscombe.
- Wootters, W. K., & Zurek, W. H. (1982). A single quantum cannot be cloned. *Nature*, 299(5886), 802–803.
- Worrall, J. (1989). Structural realism: The best of both worlds? *Dialectica*, 43, 99–124.
- Wu, T., & Tegmark, M. (2019). Toward an artificial intelligence physicist for unsupervised learning. *Physical Review E*, 100, 033311.
- Wu, X., Ma, Y., Yang, J., Chen, T., & Lin, J. (2019). On the security of TRNGs based on multiple ring oscillators. In *International conference on security and privacy in communication systems* (pp. 89–107). Springer.
- Xia, Z. (1992). The existence of noncollision singularities in Newtonian systems. *Annals of Mathematics*, 135(3), 411–468.
- Yates, D. (2016). Demystifying emergence. *Ergo*, 3, 809–841.
- Zhang, W., van Leent, T., Redeker, K., Garthoff, R., Schwonnek, R., Fertig, F., Eppelt, S., Rosenfeld, W., Scarani, V., Lim, C. C. W., & Weinfurter, H. (2022). A device-independent quantum key distribution system for distant users. *Nature*, 607(7920), 687–691.
- Żukowski, M., & Markiewicz, M. (2021). Physics and metaphysics of Wigner's friends: Even performed premeasurements have no results. *Physical Review Letters*, 126, 130402.
- Zwenger, M., Briegel, H. J., & Dür, W. (2016). Measurement-based quantum communication. *Applied Physics B*, 122(3), 50.

Name Index

A

Abellán, C., 157, 173
Acín, A., 157, 173, 190
Adam, C., 227
Adelberger, E.G., 128
Agresti, I., 190
Alarcón, A., 157, 173
Alba, E., 286
Alberts, B., 41
Alice, 50, 166–173, 176–179, 190, 191, 208, 210, 287–290
Alicki, R., 198
Allen, L., 298
Allert, R.D., 292
Allman, M.S., 157
Amant, R.S., 342
Amaya, W., 157
Amit, O., 147
Ampère, A.-M., 239
Andersen, C.K., 286
Andersen, U.L., 298
Anderson, P.W., 31, 33
Anscombe, G.E.M., 32, 34, 73, 159, 327
Antonoglou, I., 57, 306
Aolita, L., 190
Apostolou, V., 313
Appleby, M., 225
Ardón, P., 342
Aristotle, 43, 56, 82, 158
Arndt, M., 141
Arrasmith, A., 313
Artacho, E., 236
Arunachalam, S., 308
Asenbeck, B.E., 141
Aslam, N., 292

Aspect, A., 157
Aspelmeyer, M., 173
Aspuru-Guzik, A., 237, 247, 305
Auletta, G., 114
Austin, J.L., 24, 56

B

Babazadeh, A., 299
Babbush, R., 237, 313
Bäckström, S., 229
Bacon, F., 6
Badurek, G., 141
Bagan, E., 144, 322
Bahcall, J.N., 192
Bahl, S., 342
Balantekin, A.B., 128
Ballentine, L.E., 106
Barends, R., 237
Barnett, S.M., 121, 304
Barreiro, J.T., 193, 313
Barto, A.G., 73, 259, 265, 275, 277–280
Bartschat, K., 39
Bates, R., 238
Bauer, P., 162
Bauspiess, W., 141
Baysan, U., 32
Beattie, C., 306
Becker, S., 235
Beijersbergen, M.W., 298
Bell, J.S., 119, 127, 129, 156, 157, 162–174, 184, 190, 191, 205–208, 281, 290, 297
Bellemare, M.G., 306
Bellman, R.E., 275, 277
Belnap, N., 20, 158, 206

- Bemmerer, D., 128
 Bengtsson, I., 175, 225
 Benjamin, S.C., 313
 Bennett, C.H., 176, 177, 288–290, 298
 Bergerault, A., 313
 Bergou, J.A., 144, 322
 Bernien, H., 157
 Bernstein, H.J., 314, 315
 Berryman, S., 154
 Bertani, P., 314, 315
 Bertulani, C.A., 128
 Beyer, J., 157
 Biamonte, J., 308
 Bienfang, J.C., 157
 Bierhorst, P., 157, 190, 191
 Bjerland, Ø.F., 280
 Black, M., 93
 Blais, A., 286
 Blatt, R., 193, 235, 237, 286, 313
 Blok, M., 157
 Blume-Kohout, R., 224
 Blümmler, P., 292
 Bluvstein, D., 30, 197
 Bob, 50, 166–173, 176–179, 190, 191, 208, 287–290
 Bohg, J., 260, 342, 351
 Bohm, D., 163, 232
 Bohr, N., 94, 100, 121, 127, 150, 223, 232, 233, 239
 Bonilla Ataides, J.P., 30
 Bonse, U., 141
 Borges, J.L., 241
 Bos, H., 189
 Bose, S., 147
 Bourennane, M., 299
 Boyajian, W.L., 280
 Boyd, R.W., 298, 299
 Braitenberg, V., 348
 Brandl, M.F., 193
 Brassard, G., 176, 177, 288–290, 298
 Bratman, M.E., 20
 Bravyi, S., 282
 Bray, A., 39
 Bremer, B.J., 247, 305
 Brentano, F., 19
 Bridgland, A., 238
 Briegel, H.J., 17, 63, 71, 74, 141, 173, 176, 180–185, 190, 197, 221, 222, 235, 257, 259, 262, 265–271, 280, 281, 286, 287, 289–291, 293, 300–308, 310, 311, 313, 314, 316, 317, 319, 321, 323–325, 361, 369, 370
 Briegel, K.D., 292
 Brock, O., 260, 342, 351
 Brooks, R.A., 364
 Browne, D.E., 181
 Brukner, Č., 173
 Brun, T.A., 200
 Brydges, T., 237
 Bucher, D.B., 292
 Büchler, H.P., 313
 Budroni, C., 173
 Busch, P., 144
 Bustamante, C.J., 30
 Butterfield, J., 32, 33, 117

C
 Cabello, A., 156
 Cain, M., 30
 Calsamiglia, J., 196, 322
 Campbell, B., 237
 Candlish, S., 228
 Capes, J., 209
 Carlson, J., 192
 Carneiro, R.L., 33
 Cartwright, N., 23
 Carvachol, G., 190
 Cavalcanti, D., 190
 Caves, C.M., 224
 Cerezo, M., 313
 Cerf, N.J., 299
 Chakraborty, A., 342
 Chaves, R., 190
 Chemla, Y.R., 30
 Chen, J.-W., 128
 Chen, L., 342
 Chen, P.-Y., 308
 Chen, S.Y.-C., 308
 Chen, T., 189
 Chen, Y., 237
 Chen, Z., 237
 Chi, X., 41
 Chiaro, B., 237
 Chiribella, G., 121
 Choi, J., 319
 Christensen, B.G., 157, 189–191
 Christensen, N., 186
 Chuang, I.L., 123, 139, 179, 184, 195, 198, 202, 204
 Chwalla, M., 193, 313
 Cincio, L., 313
 Cirac, J.I., 190, 197, 289, 290
 Clancey, W.J., 365
 Clark, A., 308
 Clarke, R., 209
 Clausen, J., 280
 Clauser, J., 157

Cleland, A.N., 283
 Clements, W.R., 314
 Cloeren, H.J., 23
 Coakley, K.J., 157
 Cohen-Tannoudji, C., 106
 Colella, R., 141
 Coles, P.J., 313
 Combes, J., 244
 Commins, E.D., 95
 Cong, I., 30
 Conway, J.H., 62, 161
 Cory, D.G., 244
 Costantini, H., 128
 Cottingham, J., 227
 Cottrell, S.S., 144
 Couder, M., 128
 Courtial, J., 304
 Coveney, P.V., 237
 Crépeau, C., 176, 177, 288, 290, 298
 Cronin, A.D., 141
 Cross, A.W., 282
 Crossan, S., 238
 Cyburt, R.H., 40, 128

D

Dalibard, J., 156
 D'Ariano, G.M., 121
 Davids, B., 128
 Davidson, D., 43, 336
 Day, D., 206
 Dayan, P., 277
 De Finetti, B., 223
 de la Giroday, A.B., 190
 De las Cuevas, G., 257, 262, 265–268, 310, 311, 361
 De Regt, H.W., 324
 de Wolf, R., 308
 DeBroda, J.B., 221, 223, 228
 Del Colle, R., 28
 Delfosse, N., 285, 287
 Democritus, 154
 Dennett, D.C., 42, 50, 62, 161
 Dennis, E., 283, 286
 Descartes, R., 227
 Deutsch, D., 289
 Di Paolo, A., 286
 Diaconis, P., 187
 Dieks, D., 200, 223
 Diestel, R., 181
 DiNardo, J., 7
 Ding, N., 237
 Diu, B., 106
 Dobkowski, O., 147

Douguet, N., 39
 Dowty, D.R., 48
 Draeger, J.C., 292
 Dréau, A., 157
 Dretske, F., 5
 Dreyfus, H.L., 11, 12, 318, 333–339, 341, 342, 350, 352–354, 363, 364, 367, 369–372
 Driesch, H., 43, 44
 Drühl, K., 143, 145
 Ducheyne, S., 186
 Dunjko, V., 141, 221, 222, 235, 280, 287, 300–308, 310, 311, 313, 316, 321
 Dunsworth, A., 237
 Dupré, J., 23, 34, 41
 Dür, W., 129, 176, 180, 182, 190, 196, 289, 290
 Dürr, D., 39
 Dürr, S., 322
 Dyer, S.D., 157

E

Eagen, C.F., 141
 Earman, J., 33, 110, 160, 223
 Ebadi, S., 30, 197
 Echeñique, J., 56
 Edmonds, J., 285
 Edmunds, C., 235
 Ehrenfest, P., 97
 Eichler, C., 286
 Einstein, A., 37, 97, 98, 100, 110, 156, 163, 164, 168, 186
 Eisert, J., 180, 182, 235
 Ekert, A., 190, 289
 Elliott, T.J., 222
 Endo, S., 313
 Engels, F., 33
 Englert, B.-G., 92, 96, 114, 124, 127, 129, 130, 143–145, 148, 152, 225
 Englund, D., 141
 Eppelt, S., 190
 Eppl, M., 75
 Erhard, A., 286
 Erhard, M., 299, 300, 302
 Estebaran, J., 156
 Eva, B., 281, 304, 307, 317, 324, 325, 370
 Evans, R., 238
 Eve, 190, 191
 Evered, S.J., 30

F

Falkenburg, B., 33
 Farr, W.H., 157

Felappi, G., 229
 Ferrero, L., 18, 334
 Ferrie, C., 235
 Fertig, F., 190
 Feynman, R.P., 140
 Fickler, R., 299, 300, 302, 304
 Fiderer, L.J., 185, 221, 308, 313, 314, 316, 317, 319, 323
 Fidjeland, A.K., 306
 Fields, B.D., 40
 Figurnov, M., 238
 Flamini, F., 221, 313, 314, 316, 317, 319, 323, 369
 Flammia, S.T., 225, 235
 Flores, F., 342
 Flynn, B., 238
 Folman, R., 147
 Forman, P., 94
 Fortescue, B., 298
 Fortunato, M., 114
 Fossa, F., 45
 Fowler, A.G., 237, 283, 285
 Franceschetto, G., 313, 317
 Franke-Arnold, S., 304
 Frankfurt, H.G., 77
 Frauchiger, D., 145, 221
 Freedman, S.J., 128
 Frege, G., 26
 French, S., 187, 222, 223, 228, 233
 Friedrich, B., 94, 97, 98
 Frigg, R., 234, 240, 324
 Friis, N., 141, 235, 286, 287, 302
 Frisch, R., 98
 Friston, K., 308
 Fuchs, C.A., 221, 223–225, 227, 231, 233, 242
 Fujii, K., 313
 Furusawa, A., 298

G

Gachechiladze, M., 173
 Gai, M., 128
 Galileo Galilei, 6
 Gallagher, S., 334
 Gallow, J.D., 31
 Galton, F., 326
 Gambetta, J.M., 282
 Ganahl, E., 293
 Gao, X., 30
 García, A., 128
 García-Alcaine, G., 156
 García-Escartín, J.C., 189
 García-Ripoll, J.J., 286

Gardiner, C.W., 197
 Garner, A.J.P., 222
 Garthoff, R., 190
 Gaspar, J.M., 313
 Gassendi, P., 159
 Gazit, D., 128
 Geim, A.A., 30
 Genois, E., 286
 Genovese, M., 164, 171, 173
 Gentile, A.A., 238
 Georgie, Y.K., 75
 Gerlach, W., 95, 97
 Gerlich, S., 141
 Gerrits, T., 157
 Gettier, E.L., 55, 56
 Gialanella, L., 128
 Gibb, S., 31
 Gibson, J.J., 333, 340–343, 345, 354
 Gisin, N., 299
 Giuffrida, C., 189
 Giustina, M., 157
 Glackin, C., 73
 Glancy, S., 157, 189–191
 Glauber, R.J., 221
 Gleason, A.M., 156
 Glesener, G., 293
 Glick, D., 226
 Goan, H.-S., 308
 Goffman, E., 81
 Goldstein, H., 114
 Goldstein, J.A., 31
 Goldstein, S., 39
 Gottesman, D., 180, 202, 298
 Goudsmit, S.A., 95, 239
 Graepel, T., 57
 Granade, C.E., 238, 244
 Grassl, M., 181, 282
 Graves, A., 306
 Gray, K., 16, 48
 Green, T., 238
 Greenberger, D.M., 144, 183, 202
 Greife, U., 128
 Greiner, M., 30, 197
 Gröblacher, S., 173
 Gross, D., 235
 Gruber, G., 94, 97
 Gu, M., 33, 222
 Guckelsberger, C., 73
 Guerini, L., 190
 Guez, A., 57
 Gühne, O., 173, 183, 235
 Gullans, M.J., 30
 Gunning, D., 319

Gustafson, S.J., 108, 124
 Gustafsson, M., 229
 Gyurik, C., 308, 313, 316

H

Hacking, I., 6, 30, 138, 155, 246
 Häffner, H., 193
 Hafner, V.V., 63, 75
 Hájek, A., 230
 Hamann, A., 141, 314
 Hamel, D.R., 157
 Handsteiner, J., 157
 Hangl, S., 290, 305
 Hangleiter, D., 30
 Harrington, P.M., 313
 Harris, N.C., 141
 Hartman, L., 206
 Hartmann, L., 196
 Hartmann, S., 234, 240, 324
 Häse, F., 247, 305
 Hasle, P., 63
 Haslinger, P., 141
 Hass, M., 128
 Hassabis, D., 57, 238, 306
 Hausman, K., 260, 342, 351
 Haxton, W.C., 128
 Hayes, D., 190
 Hayes, W.B., 159
 Hazeleger, W., 162
 Head-Gordon, M., 236
 Healey, R., 91, 117, 223
 Heeger, K., 128
 Hegel, G.W.F., 33, 43
 Heidegger, M., 11, 50, 333–339, 341–354, 357, 361, 363, 366, 367, 369
 Heil, J., 31
 Hein, M., 179–182
 Heisenberg, W., 95, 108, 111, 114, 126, 152
 Hellings, C., 286
 Hempel, C., 237
 Henkel, C., 147
 Hennrich, M., 193, 286, 313
 Hensen, B., 157
 Hentschel, K., 94
 Herbert, N., 200
 Herrero-Collantes, M., 189
 Herrmann, J., 286
 Herschbach, D., 98
 Hersh, R., 81
 Hiley, B.J., 232
 Hilgevoord, J., 150
 Hillery, M., 144, 322
 Hoang, M.C., 225

Hochberg, M., 141
 Hochrainer, A., 157
 Hodge, C., 157
 Hofmann, C., 39
 Hollenberg, L.C.L., 285
 Holmes, S., 187
 Horne, M.A., 183, 202
 Horodecki, K., 175
 Horodecki, M., 175
 Horodecki, P., 175
 Horodecki, R., 175
 Horton, T.E., 342
 Horthy, J., 45
 Howard, D., 32
 Howes, A., 276
 Huber, M., 235, 299, 300, 302
 Hubert, T., 57
 Humberstone, L., 20
 Hume, D., 28
 Humphreys, P., 32, 33
 Humphreys, P.C., 314
 Hurwitz, R., 139
 Husserl, E., 338
 Hyllus, P., 173, 183

I

Ichikawa, J.J., 55
 Imbriani, G., 128
 Ismael, J., 102
 Itahashi, T., 128

J

Jain, U., 342
 Jakobsen, D., 63
 Jaksch, D., 197
 James, W., 83
 Janzing, D., 28
 Japha, Y., 147
 Jasra, A., 235
 Jeffrey, E., 237
 Jennewein, T., 157
 Jerbi, S., 185, 221, 306, 308, 313, 314, 316, 321
 Johnson, A., 41
 Jones, D.T., 238
 Joo, J., 286
 Jordan, S., 189–191
 Joshi, M.K., 237
 Jozsa, R., 177, 288–290, 298
 Jumper, J., 238
 Junghans, A., 128
 Jurcevic, P., 237

K

Kaelbling, L.P., 265, 277
 Kalb, N., 157
 Kalinowski, M., 30, 197
 Kaltenbaek, R., 173
 Kamerlingh Onnes, H., 246
 Kane, R., 230
 Kant, I., 345
 Kaplan, F., 73
 Karlsson, A., 299
 Karolyshyn, T., 30
 Kavukcuoglu, K., 238, 306
 Kayhan, E., 75
 Keesling, A., 197
 Keet, A., 298
 Keil, G., 77
 Kelly, J., 237
 Kempe, J., 310
 Kepler, J., 159
 Kes, P., 246
 Kewming, M.J., 222
 Kim, J., 31
 King, H., 306
 Kirkpatrick, J., 238
 Kitaev, A., 283, 286, 298
 Kivlichan, I.D., 237
 Klappenecker, A., 181
 Knauer, S., 238
 Knill, E., 157, 189–191
 Knuuttila, T., 240
 Koch, W., 39
 Kochen, S., 156
 Kofler, J., 143, 157
 Kohli, P., 238
 Kokail, C., 237
 Kolthammer, W.S., 314
 Kragic, D., 260, 342, 351
 Krenn, M., 299–304, 306
 Krinner, S., 286
 Krumm, M., 185, 221, 307, 308, 313, 314, 316, 317, 319, 323
 Kübler, J.M., 221
 Kubodera, K., 128
 Kueng, R., 235
 Kuhn, T.S., 24, 92
 Kumaran, D., 57, 306
 Kumor, D.R., 157
 Kusch, P., 102, 139
 Kwiat, P.G., 157

L

Lacroix, N., 286
 Ladyman, J., 25, 232

Lahti, P.J., 144
 Lai, M., 57
 Laing, A., 141, 238
 Laloë, F., 106
 Lambrocco, C., 157
 Lanctot, M., 57
 Landahl, A., 283, 286
 Landé, A., 94, 239
 Landsman, K., 33
 Langanke, K., 128
 Langford, J., 265
 Lanyon, B.P., 237
 Lapkiewicz, R., 299, 304
 Laplace, P.S., 160
 Larmer, R., 28
 Larsson, J.-A., 157
 Laudan, L., 155, 232
 Law, K.J.H., 235
 Lawson, R., 229
 Lazar, S., 286
 Leach, J., 304
 Lear, J., 2
 LeBlanc, F., 38
 Legg, S., 306
 Leighton, R.B., 140
 Leitner, D., 128
 Leitner, M., 128
 LeMaitre, P.A., 307, 317
 Lendi, K., 198
 Leroux, C., 286
 Lesanovsky, I., 313
 Levine, H., 197
 Levine, M.P., 28
 Lewes, G.H., 32
 Lewis, D.K., 64, 158, 161
 Lewis, J., 41
 Lewis, R.L., 276
 Li, S.H., 30
 Li, Y., 41
 Libbrecht, K.G., 187
 Lidar, D.A., 200
 Lillicrap, T., 57
 Lim, C.C.W., 190
 Lin, J., 189
 Lindblad, G., 198
 Linden, N., 177
 Linson, A., 308
 Lita, A.E., 157
 Littman, M.L., 277
 Liu, S., 30
 Liu, Y.-K., 189–191, 235
 Lloyd, S., 308
 Lockwood, M., 127
 Lohan, K.S., 342

López-Incera, A., 305, 369
 Lord Kelvin, 92
 Lorentz, H.A., 94, 239
 Lougovski, P., 235
 Love, P.J., 228, 237
 Loviken, P., 63
 Löwe, B., 17
 Lucero, E., 237
 Lüdde, H.J., 97, 98
 Lukens, J.M., 235
 Lukin, M.D., 30, 197, 292
 Luo, L., 190

M

Ma, X., 308
 Ma, X.-S., 143
 Ma, Y., 189
 Macchiavello, C., 289
 Mackenzie, D., 7, 28, 49, 187
 Mackie, J.L., 26
 Magniez, F., 310, 311
 Maier, C., 237
 Mair, A., 298
 Majumder, A., 185
 Ma-Kellams, C., 48
 Makmal, A., 235, 266, 280, 306, 307, 310, 311
 Malik, M., 235, 298–300, 302, 304
 Mancusi, M., 190
 Mannalath, V., 189
 Manning, T.A., 190
 Manovitz, T., 30
 Manzano, D., 266, 280
 Marcucci, L.E., 128
 Margalit, Y., 147
 Mariantoni, M., 283
 Markham, D., 298
 Markiewicz, M., 145, 221
 Marov, M.Y., 35
 Marshall, S., 308, 313, 316
 Marsili, F., 157
 Martin-Delgado, M.A., 235, 286, 310, 311
 Martinez, E.A., 193, 286
 Martinis, J.M., 283
 Marx, K., 33
 Maskara, N., 30, 197
 Maslov, D., 282
 Massar, S., 190
 Matsukevich, D.N., 190
 Maudlin, T., 39
 Maunz, P., 190
 Mautner, J., 266, 280
 May, R.M., 192
 Mazumdar, A., 147
 McClean, J.R., 237, 313
 McConnell, G., 225
 McDowell, J.H., 369–372
 McGrew, T., 28
 McKay, B., 156
 McGill, N.D., 156
 Megrant, A., 237
 Meincke, A.S., 32
 Melnikov, A.A., 280, 290, 300–304, 306, 307
 Melville, H., 17
 Mendonca, R., 342
 Merleau-Ponty, M., 352
 Merlet, J.-P., 156
 Mermin, N.D., 171, 174, 202, 223, 225, 227, 233, 242
 Merz, M., 240
 Metcalf, B.J., 314
 Meth, M., 235, 286
 Meyer, R., 186
 Meyer-Scott, E., 157
 Michelson, A.A., 92
 Migdall, A.L., 157
 Milburn, A., 189
 Milburn, G.J., 222
 Milgram, E., 230
 Miller, J., 185
 Miller, T., 319
 Millman, S., 102, 139
 Mink, A., 189–191
 Minsky, M., 279, 334
 Mirin, R.P., 157
 Mishra, S., 189
 Mitarai, K., 313
 Mitchell, K.J., 18, 376
 Mitchell, M.W., 157
 Mittelstaedt, P., 144
 Miyake, A., 185
 Mnih, V., 306
 Molina-Terriza, G., 298, 299
 Monroe, C., 190
 Montague, M., 229
 Montgomery, R., 187
 Monz, T., 193, 235, 237, 286, 313
 Moore, A.W., 277
 Morales, J., 209
 Moretti, L., 55
 Morgan, D., 41
 Morgan, M.S., 241
 Morrison, M., 33, 241
 Motobayashi, T., 128
 Moukouri, D., 147
 Moukouri, S., 147
 Mueller, E.J., 313
 Mukhamedzhanov, A., 128

Mulder, J.M., 27, 32, 34, 43, 72
 Müller, J., 209
 Müller, M., 286, 313
 Müller, T., 17, 27, 63, 64, 71, 74, 156, 158,
 207, 209, 221, 269–271, 280, 281, 304,
 305, 307, 313, 314, 316, 317, 319,
 323–325, 370
 Murch, K.W., 313
 Murdoch, D., 227
 Muschik, C.A., 237
 Mutus, J.Y., 237
 Myrvold, W., 91, 154, 164, 171, 173

N

Nagel, T., 1, 2, 22, 42, 55, 60–62, 206, 376
 Nagy, A., 30
 Nam, S.W., 157, 189–191
 Napolitano, J., 96, 100, 106, 171
 Nayak, A., 310, 311
 Nebendahl, V., 193
 Neeley, M., 237
 Neergaard-Nielsen, J.S., 298
 Negnevitsky, V., 286
 Neill, C., 237
 Nelson, A.W.R., 238
 Neuman, K.C., 30
 Neven, H., 237
 Newton, I., 129, 159, 186
 Nguyen, P.D.H., 75
 Ni, H., 39
 Nicholson, D.J., 34
 Nickerson, N.H., 285
 Nielsen, M.A., 33, 123, 179, 184, 195, 198,
 202, 204
 Nigg, D., 193, 286, 313
 Nimmrichter, S., 141
 Noble, D., 41
 Noble, R., 41
 Nollett, K.M., 128
 Norman, D.A., 342
 Norris, G.J., 286
 Norton, J.D., 110
 Norvig, P., 17, 43, 185, 256, 257, 259
 Nouroozi, R., 299
 Nouvian, M., 305
 Nunes, F.M., 128

O

Obenchain, T.G., 155
 Occam, 23, 24
 O'Connor, T., 32
 Odegard, V.H., 41

Øhrstrøm, P., 63
 Olive, K.A., 40
 Olmschenk, S., 190
 Olshanii, M., 223
 O'Malley, P.J.J., 237
 Ometto, D., 27
 Oreskes, N., 78, 241
 Ørsted, C., 239
 Orsucci, D., 311
 Ostrovski, G., 306
 Oudeyer, P.-Y., 73
 Overhauser, A.W., 141

P

Padgett, M.J., 304
 Paesani, S., 238
 Pairat, È., 342
 Pancotti, N., 308
 Paparo, G.D., 235, 310, 311
 Paraniak, M.M., 148
 Paris, M.G.A., 235
 Parisi, G., 114
 Park, H., 292
 Park, T.-S., 128
 Parker, P.D., 128
 Pascual, P., 108
 Paterek, T., 173
 Pathak, A., 189
 Pathak, D., 342
 Pauli, W., 95
 Pavičić, M., 156
 Pazem, J., 308
 Pearl, J., 7, 28, 49, 50, 236
 Pechukas, P., 198
 Penedones, H., 238
 Perales, A., 33
 Pereboom, D., 28, 61, 64
 Peres, A., 123, 145, 156, 176, 177, 288, 290,
 298
 Perinotti, P., 121
 Perloff, M., 20
 Pesic, P., 6
 Peters, J., 28
 Petersen, S., 238, 306
 Petrick, R.P.A., 342
 Petruccione, F., 308
 Pettigrew, R., 230
 Phelan, K., 157
 Phipps, T.E., 98
 Piater, J., 290, 305
 Piazza, T., 55
 Pichler, H., 197
 Pickard, H., 24

Pico Villalpando, A., [63](#)
 Pienaar, J., [235](#), [236](#)
 Pienaar, J.L., [223](#)
 Pironio, S., [190](#)
 Placek, T., [63](#), [156](#), [158](#), [206](#), [207](#)
 Plato, [25](#)
 Poderini, D., [190](#)
 Podolsky, B., [156](#), [164](#), [168](#)
 Polani, D., [73](#)
 Pombo, O., [24](#)
 Poole, C.P., [114](#)
 Popescu, S., [177](#), [289](#)
 Postler, L., [235](#), [286](#)
 Potapenko, A., [238](#)
 Poulsen Nautrup, H., [185](#), [221](#), [286](#), [287](#),
[300–304](#), [306](#), [314](#), [321](#), [369](#)
 Prati, P., [128](#)
 Preskill, J., [282](#), [283](#), [286](#), [298](#)
 Pritchard, D.E., [141](#)
 Pritzel, A., [238](#)
 Pruneri, V., [157](#)
 Putnam, H., [206](#)

Q

Qi, J., [308](#)
 Qin, C., [238](#)
 Qiu, X., [41](#)
 Quine, W.V.O., [22](#), [23](#)
 Quint, S., [193](#)
 Quintana, C., [237](#)

R

Rabi, I.I., [102](#), [139](#)
 Radkohl, T., [185](#)
 Raff, M., [41](#)
 Ragab, H., [189](#)
 Rall, P., [282](#)
 Ramamoorthy, S., [308](#), [342](#)
 Ramsey-Musolf, M.J., [128](#)
 Rapp, J.T., [247](#), [305](#)
 Rarity, J.G., [238](#)
 Rauch, H., [141](#)
 Raussendorf, R., [180–182](#), [184](#), [185](#), [197](#), [291](#)
 Razavi, K., [189](#)
 Rebentrost, P., [308](#)
 Reck, M., [314](#), [315](#)
 Redeker, K., [190](#)
 Řeháček, J., [235](#)
 Reiserer, A., [157](#)
 Remm, A., [286](#)
 Renes, J.M., [224](#)
 Renner, R., [145](#), [221](#)

Rheinberger, H.-J., [235](#), [246](#)
 Ricou, A., [313](#), [317](#)
 Ried, K., [281](#), [304](#), [305](#), [307](#), [317](#), [324](#), [325](#),
[370](#)
 Riedmiller, M., [306](#)
 Rietdijk, C., [206](#)
 Ringbauer, M., [235](#), [286](#)
 Riska, D.O., [192](#)
 Robb, D., [31](#)
 Roberts, K., [41](#)
 Robertson, R.G.H., [128](#)
 Roch, L.M., [247](#), [305](#)
 Roger, G., [156](#)
 Roland, J., [310](#), [311](#)
 Romero, J., [237](#)
 Romero, P.A., [247](#), [305](#)
 Rommal, A., [189–191](#)
 Ronneberger, O., [238](#)
 Roos, C.F., [193](#), [237](#), [313](#)
 Rosen, N., [156](#), [164](#), [168](#)
 Rosenfeld, W., [190](#)
 Ross, D., [25](#)
 Rötteler, M., [181](#)
 Roushan, P., [237](#)
 Ruitenberg, J., [157](#)
 Rumberg, A., [158](#)
 Russell, B., [17](#), [26–28](#), [56](#)
 Russell, S., [17](#), [43](#), [185](#), [256](#), [257](#), [259](#)
 Russo, F., [79](#), [81](#)
 Rusu, A.A., [306](#)
 Ryle, G., [242](#)

S

Sadik, A., [306](#)
 Saffman, M., [195](#)
 Saffko, J.L., [114](#)
 Saggio, V., [141](#)
 Sakurai, J.J., [96](#), [106](#), [106](#), [171](#)
 Sales Rodriguez, P., [30](#)
 Salge, C., [73](#)
 Sanders, B.C., [298](#)
 Sanders, S., [185](#)
 Sands, M.I., [140](#)
 Sank, D., [237](#)
 Sankaran, B., [260](#), [342](#), [351](#)
 Sanpera, A., [289](#)
 Santagati, R., [238](#)
 Santha, M., [310](#), [311](#)
 Sartenaer, O., [32](#)
 Sartre, J.-P., [76](#)
 Sauer, T., [97](#), [98](#)
 Scarani, V., [190](#)
 Schaal, S., [260](#), [342](#), [351](#)

Schack, R., 221, 223–225, 227, 233, 242
 Schatz, D.G., 41
 Schear, J.K., 369
 Scheidl, T., 157
 Schiansky, P., 141
 Schiavilla, R., 128, 192
 Schillaci, G., 63
 Schindler, P., 193, 235, 286, 313
 Schlingemann, D., 181
 Schmidt, L., 97, 98
 Schmidt-Böcking, H., 94, 97, 98
 Schmiedmayer, J., 141
 Schölkopf, B., 28
 Schouten, R., 157
 Schrittwieser, J., 57
 Schrödinger, E., 95, 111–118, 123, 126–131, 164, 220, 221, 223, 294, 316
 Schuld, M., 308
 Schumacher, B., 289
 Schwinger, J., 96, 118, 130, 151, 152
 Schwonnek, R., 190
 Sciarino, F., 141, 190
 Scott, A.J., 224
 Scully, M.O., 130, 143, 145
 Searle, J., 336, 338
 Segré, E., 98
 Sekatski, P., 129
 Semeghini, G., 30, 197
 Senior, A.W., 238
 Shalm, L.K., 157
 Shaw, M.D., 157
 Shen, H., 237
 Shimony, A., 164, 171, 173, 183, 202
 Shrapnel, S., 222
 Shvetsov-Shilovski, N.I., 39
 Sifre, L., 57, 238
 Sigal, I.M., 108, 124
 Silver, D., 57, 238, 306
 Silvi, P., 237
 Simonyan, K., 57, 238
 Simpson, E.C., 128
 Sinayskiy, I., 308
 Singh, S., 276
 Skotiniotis, M., 129
 Smith, I.D., 185
 Smolin, J.A., 289
 Snover, K.A., 128
 Sourjik, V., 60
 Specker, E., 156
 Spiller, T.P., 286
 Spitaleri, C., 128
 Spreeuw, R.J.C., 298
 Stacey, B.C., 223, 225

Stadler, M., 286
 Stefik, M., 319
 Steinlechner, F., 157
 Stern, J.A., 157
 Stern, O., 93, 95–103, 107, 109, 110, 113, 117, 119–120, 126, 128–149, 151–153, 155, 156, 164, 166, 167, 199, 210, 242, 243, 291
 Steup, M., 55
 Stevens, B., 162
 Stevens, M.J., 157
 Steward, H., 2, 8, 18–21, 50, 62, 63, 65, 66, 68, 69, 71, 72, 74, 158
 Stoothoff, R., 227
 Straub, I., 59
 Strawson, P.F., 17, 34, 52, 53
 Stricker, R., 235, 286
 Strieder, F., 128
 Strömberg, T., 141
 Stumpf, S., 319
 Suemmerer, K., 128
 Sukhatme, G.S., 260, 342, 351
 Sutton, R.S., 259, 265, 275, 277–280
 Swenson, P., 209
 Swiadek, F., 286
 Szedmak, S., 290
 Szegedy, M., 310

T

Tannery, P., 227
 Taylor, C., 363, 367, 370–372
 Taylor, J.M., 222, 308
 Tegmark, M., 45
 Templeton, A., 97, 98
 Tennant, N., 26
 Terhal, B.M., 200
 Thomas, L., 95
 Thomason, R.H., 45
 Thompson, J., 222
 Thompson, M., 44, 66, 344
 Thompson, M.G., 141
 t'Hooft, G., 39, 161, 207, 232
 Thorne, K.S., 186
 Tiersch, M., 266, 280, 293, 300–304, 306
 Timpson, C.G., 226
 Tomasello, M., 18, 376
 Torner, L., 298, 299
 Torres, J.P., 298, 299
 Tortorici, E., 157
 Tóth, G., 173, 183, 235
 Trageser, W., 97, 98
 Tranter, A., 237

Trautvetter, H.-P., 128
 Trenkwalder, L.M., 280, 306, 314, 321, 369
 Tribble, R.E., 128
 Tulving, E., 258
 Tumulka, R., 39
 Tunyasuvunakool, K., 238
 Turek, M., 319
 Turing, A.M., 5, 335
 Turner, M.J., 292
 Typel, S., 128

U

Uberseder, E., 128
 Uebel, T.E., 24, 227
 Uffink, J., 150
 Ugur, E., 290
 Uhlenbeck, G.E., 95, 239
 Urbach, E.K., 292
 Ursin, R., 157

V

Vaidman, L., 39, 232
 Vainsencher, A., 237
 van Bijnen, R., 237
 Van Delft, D., 246
 Van den Nest, M., 180, 182
 Van Gulick, R., 26
 Van Helden, A., 159
 Van Inwagen, P., 63
 van Leent, T., 190
 van Leeuwenhoek, 69
 Van Lier, M., 66, 247
 van Loock, P., 298
 Van Miltenburg, N., 27
 Van Otterlo, M., 274–277
 Van Riel, R., 26, 63
 Vandersypen, L.M.K., 139
 Vaziri, A., 298
 Velleman, J.D., 2, 60
 Veness, J., 306
 Verma, V.B., 157
 Vermeulen, R., 157
 Versteegh, M.A.M., 157
 Vetter, P., 128
 Vidal, G., 177
 Vineberg, S., 230
 Vining, A.Q., 308
 Vitagliano, G., 235, 302
 Vladimirov, N., 60

von Grafenstein, N.R., 292
 Von Neumann, J., 108, 156
 Vorm, E., 319
 Vuletić, V., 30, 197

W

Wallnöfer, J., 290
 Wallraff, A., 286
 Walmsley, I.A., 314
 Walsworth, R.L., 292
 Walter, P., 41
 Walther, H., 143, 145
 Walther, P., 141
 Wang, F., 299
 Wang, J., 141
 Wang, J.Y., 319
 Wang, M.D., 30
 Wang, S.X., 193
 Wang, T.T., 197
 Watkins, C.J.C.H., 277
 Watson, G., 52
 Wayne, M.A., 157
 Weedbrook, C., 33
 Wegner, D., 16, 48, 64
 Weihs, G., 298
 Weimer, H., 313
 Weinfurter, H., 190
 Weiss, M.B., 223
 Wellmann, J., 69
 Wengerowsky, S., 157
 Wenner, J., 237
 Wermter, S., 75
 Werndl, C., 160
 Werner, R.F., 289
 Werner, S.A., 141
 Wheeler, J.A., 145, 233
 Wheeler, M., 11, 335, 370
 White, T.C., 237
 Whiteside, A.C., 285
 Wiebe, N., 238
 Wiering, M., 274–277
 Wierstra, D., 306
 Wiescher, M., 128
 Wiesner, S.J., 176, 177
 Wilfing, A., 141
 Winograd, T., 342
 Winslow, L., 128
 Wiringa, R.B., 192
 Wittek, P., 308
 Wittgenstein, L., 227, 350
 Wittmann, B., 157

Woerdman, J.P., [298](#)
Wölk, S., [141](#)
Wootters, W.K., [176](#), [177](#), [200](#), [288–290](#), [298](#)
Worrall, J., [232](#)
Wrisley, G., [228](#)
Wu, T., [45](#)
Wu, X., [189](#)

X

Xia, Z., [160](#)
Xu, M., [20](#)

Y

Yang, C.-H.H., [308](#)
Yang, G.-Z., [319](#)
Yang, J., [189](#)
Yard, J., [225](#)
Yasin, A., [144](#)
Yates, D., [32](#)
Yeh, T.-H., [40](#)

Yoder, T.J., [282](#)
Yuan, X., [313](#)

Z

Zacharias, J.R., [102](#), [139](#)
Zahavi, D., [334](#)
Zanghi, N., [39](#)
Zeeman, P., [93–95](#), [238](#), [325](#)
Zeilinger, A., [141](#), [143](#), [157](#), [173](#), [183](#), [202](#),
[298–304](#), [306](#), [314](#), [315](#)
Zhang, T., [265](#)
Zhang, W., [190](#)
Zhang, Y., [157](#), [189–191](#)
Zhou, H., [30](#), [292](#)
Zhou, Z., [147](#)
Žídek, A., [238](#)
Zoller, P., [190](#), [197](#), [237](#), [289](#), [290](#), [313](#)
Żukowski, M., [145](#), [173](#), [221](#)
Zurek, W.H., [145](#), [200](#), [233](#)
Zwerger, M., [289](#)
Życzkowski, K., [158](#), [175](#)

Subject Index

A

- Accident, 16, 28, 40, 49, 74, 95, 169, 347, 348
- Accuracy argument, 230
- Act, 1, 3, 66, 67, 138, 145, 179, 189, 196, 257, 283, 290, 292, 314, 327, 356
- Action, 1, 16–18, 20, 21, 37, 43, 49, 51, 52, 59, 67, 74, 76, 80–82, 163, 168, 174, 175, 180, 181, 185, 186, 197, 207, 226, 230, 232, 234, 249, 257–280, 283, 286, 287, 290, 292–295, 297, 300–307, 311–313, 315, 317–320, 322–324, 327, 336, 337, 342, 343, 345, 346, 350–352, 354, 356–362, 364–366, 368–371
 - reason for, 1, 320, 371
- Activity, 45, 48, 79, 83, 247, 248, 338, 342, 346, 350, 361, 362
- Actuality, 158, 344, 345
- Adaptation, 10, 18, 71, 74, 187, 248, 258, 261, 267, 290, 327, 345–349, 360, 363
- Adaptivity, 74, 248, 343, 351
- Affordances, 11, 340–343, 345, 347, 354, 355, 357, 368
- Agency
 - animal, 5, 8, 18, 51, 333, 335, 336, 355, 376, 378
 - artificial, 5, 11, 12, 15, 18, 22, 44, 45, 55, 65–70, 75, 78, 79, 222, 239, 247, 269, 281, 305, 317, 318, 333–335, 348, 366, 378
 - attribution, 16, 21, 22, 48–71, 74, 79, 80
 - biological forms of, 15, 347
 - category of, 8, 42, 44, 45, 49
 - of cats, 4, 42, 51–54, 58, 64, 70, 220
 - constraining, 77
 - core characteristics of, 4
 - criteria of, 18, 48, 71–75, 255, 327
 - detached, 370–372
 - diachronic aspects of, 18, 73–75, 327
 - embedded, 369–372
 - exhibit, 5, 239, 335, 368
 - explication of, 8, 21, 22, 42, 43, 71–79, 207, 248, 256, 335, 336, 344, 377
 - free, 28, 59, 61, 63–65, 75–77, 346, 371
 - full-blooded, 18, 324
 - human, 3, 5–7, 11, 15, 27, 50–52, 65, 73, 82, 208, 234, 239, 333–337, 341, 355, 357, 358, 365–366, 370, 376
 - incompatibilism, 20, 21, 62, 72–73, 327
 - key aspects of, 3, 5, 256, 259, 342, 355
 - modelling, 3–6, 8, 10, 15, 18, 22, 44, 63, 75, 79–81, 220, 222, 234–249, 255, 281, 305, 306, 324–326, 334, 335, 344, 346, 349, 369, 378
 - model of, 247–248
 - moral, 3, 15, 18, 27, 28, 44, 50, 52, 53, 56, 59, 64, 76, 77
 - of non-biological systems, 22, 65, 68, 78
 - non-rigidity of, 74, 346
 - notion of, 2–4, 15–22, 27–29, 42–44, 48, 50, 51, 54, 55, 61, 62, 64, 71, 76, 78, 206, 221, 222, 230, 232, 241, 281, 308, 327, 333, 335, 365, 372, 376
 - phenomenological approach to, 5, 11, 12, 23, 50, 73, 334–336, 356, 369
 - problem of, 1–3, 8, 15, 22, 79, 219, 309, 324, 336, 375, 376
 - role of, 7, 8, 22, 79, 81, 83, 305
 - scepticism, 7, 9, 28, 39, 54–65, 67, 91, 190, 206, 207, 232, 375, 376

- Agency (*cont.*)
 scientific, 10, 18, 64, 78, 79, 220, 222, 227,
 228, 234–249, 259, 324, 325, 327, 333,
 334, 365–372, 375
 simpler forms of, 3, 42, 355
 solitary, 78
 supernatural conception of, 15, 48, 375
 suspension of, 77
- Agent
 animal, 43, 47, 58, 75, 342, 355, 366, 378
 artificial, 5, 15, 18, 22, 65–70, 75, 78, 205,
 222, 239, 247, 282, 287, 317, 319, 348,
 364–366
 autonomous, 1, 241, 304
 biological, 57, 58, 73, 78, 347
 causation, 31, 50, 52, 209
 description of, 117, 175, 220, 222, 243,
 245, 247, 260, 261, 274, 280, 309, 314,
 346
 experience of, 1, 11, 74, 80, 221, 223, 225,
 227, 228, 230, 231, 240, 242, 258, 259,
 280, 312, 320, 324, 326, 351, 352, 358,
 359
 historicity of, 18, 71, 73, 256, 327, 343
 individual development of, 229
 internal structure of, 220, 230, 378
 learning, 6, 8, 10, 11, 19, 71, 73, 75, 78,
 80, 185, 205, 220–223, 228–230, 235,
 238, 244, 245, 248–249, 255, 257–261,
 263–265, 267, 268, 272, 274–281, 287,
 290, 295–297, 302–306, 308–310, 314,
 317, 323–327, 340, 343, 348, 351, 353,
 354, 358–362, 365, 370, 377
 linguistic, 17, 43, 45, 48, 347
 memory of, 10, 19, 21, 73, 74, 221, 247,
 248, 255–258, 261, 262, 267, 269, 271,
 272, 280, 295, 302, 304, 324, 326, 339,
 351, 352, 357, 368, 369
 moral, 15, 17, 44, 53, 59, 63, 75, 76
 non-social, 78
 optimal policy of, 10, 275, 277, 279, 357
 organisation of, 3, 9, 78
 percept-action, 17, 263, 305, 309, 358
 planning, 257, 357
 reflex, 17, 185, 287, 375
 search, 257
 software, 17
- Agentive powers, 76
- Aggregate, 2, 21, 23, 27, 30, 35, 39, 40, 131,
 161–163
- Agreement, 4, 48, 62, 78, 133, 228, 349
- Animal, 3, 5, 8, 15–19, 43, 44, 47, 51, 58,
 60–62, 67, 69, 74–76, 78, 120, 333,
 335, 336, 339–343, 345, 347, 350, 355,
 366, 376–378
 higher (*see* Higher animal)
 possibilities for, 3, 5, 16, 333, 335, 336,
 340, 347
- Argument, 21, 23, 25, 28–31, 36, 38, 59–65,
 115, 130, 145, 150, 156, 164, 182, 186,
 192, 195, 206, 208, 223, 227, 230, 232,
 266, 270, 276, 278, 315, 335, 348, 350,
 355, 370
 accuracy (*see* Accuracy argument)
 cogito (*see* Cogito argument)
 consequent (*see* Consequent argument)
 developmental, 348
 Dutch book (*see* Dutch book argument)
 sceptical (*see* Sceptical argument)
 Van Inwagen's replay (*see* Van Inwagen's
 replay argument)
- Artefact, 16, 35, 44, 46, 47, 55, 65, 66, 68, 69,
 79, 157, 319, 335, 336, 339, 341, 342,
 354
- Artificial intelligence (AI), 11, 12, 28, 45, 49,
 79, 239, 255–257, 259, 269, 317–320,
 333–336, 344, 366, 372, 376, 377
 explainable (XAI), 255, 269, 318, 319
- Artificial learning, 57, 205, 317
- Artificial systems, 60, 66–68, 247, 281, 378
- Atomic spectra, 94–95, 114, 239
- Atoms, 24, 30, 40, 94, 154, 220, 287
- Attitude, 2, 3, 20, 22, 47, 50, 52, 53, 59, 66,
 149, 154, 155, 321, 354
 propositional, 20, 47
 reactive, 52
 sceptical, 2
 straight-face, 3
- Attribution, 16, 21, 22, 48–71, 74, 79, 80, 109,
 111, 236, 321, 366, 377
 agency (*see* Agency, attribution)
 causal (*see* Causal attribution)
 epistemic, 69
 ladder of, 21, 48–52, 54, 80, 236, 321
 moral, 48–54, 56, 59, 321
 ontological, 21–42, 49–51, 68–71, 80, 377
- Authenticity, 349, 367
- Automated scientific discoveries, 222
- Autonomy, 1, 15, 28, 30, 33, 43, 57, 67, 68,
 241, 302, 304
- Awareness, 227, 336, 338, 340, 342, 347,
 362
- B**
- Bacteria, 17, 47, 59, 60, 62, 69, 70, 72, 75

- Basis, 5, 8, 9, 25, 38, 42, 66, 74, 75, 93, 94,
 104–108, 111, 115, 116, 118, 121, 125,
 126, 132, 135, 141, 156, 157, 164, 165,
 168, 175, 176, 180, 181, 183, 193, 194,
 196, 200–202, 205, 207–211, 236, 259,
 290, 292, 294, 322, 334, 345, 368, 371,
 375
 Bayesian learning, 223, 235, 296, 297
 Bayesian probability theory, 223, 224, 228,
 235, 243, 244, 296
 Bayesian update, 117, 223, 228, 235, 243, 244,
 296
 Behaviour, 5, 33, 53, 54, 57, 58, 60, 64, 67–70,
 73, 74, 94, 116, 119, 123, 132, 135,
 154, 160, 244, 258, 261, 263, 278, 279,
 281, 285, 291, 296, 298, 318–321, 324,
 325, 327, 347, 348, 357, 363, 377, 378
 adaption of, 74, 258, 261, 263
 agential, 244, 258, 261, 263, 296, 324
 explanation of (*see* Explanation, of
 behaviour)
 inflexible, 57, 74
 mechanical, 68
 non-rigid, 73, 74
 Being-in-the-world, 11, 333, 335–344, 347,
 348, 352, 353, 356, 362, 363, 365–367,
 369
 as acting, 11, 336–343
 Belief, 4, 20, 48, 55, 56, 60, 159, 221, 224,
 225, 227–230, 238, 326, 336, 377
 religious, 48, 159
 state of, 4, 221, 326
 in the supernatural, 48
 update of, 225, 228, 377
 Bell basis, 175
 Bellman equations, 277
 Bell's inequality, 172, 183, 190, 191
 Bell singlet state, 164–166, 174, 175, 191, 208,
 281
 Bell's theorem, 156, 162–174, 184, 190,
 205–207, 290
 Bell tests, 119, 157, 191
 Bell type state, 174–186, 281, 298
 Biased, 59, 135, 210, 243, 279, 296
 Binding energies, 236–239
 Bitmap, 30, 158, 161–163, 207
 Black box, 318, 325
 Bohr's model of the atom, 94, 368
 Born rule, 105, 115, 118–131, 134, 150, 153,
 155, 157, 160, 167, 189, 190, 192, 209,
 210, 224, 225, 228, 234, 244, 295
 Bose-Einstein condensation, 154, 187
 Bra, 97, 103–105, 165
 Breakdown, 39, 74, 341, 342, 347, 367
 Building, 5, 16, 18, 29, 35, 46–48, 60, 66, 68,
 73, 74, 154, 175, 206, 210, 220, 225,
 226, 236, 246–248, 255, 283, 291–297,
 307, 317, 324–326, 333, 338, 362, 377
C
 Capacity for interference, 143, 197, 322
 Care, 50, 78, 149, 320, 340, 347, 349, 355,
 366–368
 Cat, 3, 4, 36, 42, 45–47, 51–54, 58, 64, 70,
 119, 220, 221, 269, 270
 agency of (*see* Agency, of cats)
 Causal attribution, 1, 49–51, 53, 74, 321
 Causal difference, 18
 Causal happening, 49
 Causal impact, 1, 50, 377
 Causal influence, 31, 168, 171
 Causal power, 53, 66, 68
 Causal slack, 62, 80
 Causation, 7, 15, 27, 28, 31–32, 34, 49–52, 54,
 80, 210, 236
 indeterministic, 32
 ladder of (*see* Ladder of causation)
 mental, 31
 notion of, 27, 28, 31, 34, 209
 top-down, 31–32, 34
 Cause and effect, 7, 32
 Channel, 98–103, 109, 110, 113, 123, 132,
 133, 137, 138, 140, 141, 146, 149, 167,
 168, 177, 195–198, 200, 201, 203, 209,
 210, 242–244, 284, 287–292
 bit-flip, 196, 198
 dephasing, 195, 197, 202
 depolarizing, 198, 288
 noise, 195–197, 199, 200, 203, 287–292
 Pauli, 196–198, 201, 202, 284, 288, 289
 phase-flip, 195, 196, 198
 Circumspection, 337–338, 340, 342, 343, 351,
 353, 355, 361–365, 368, 370, 377–378
 Classical dynamics, 92, 94, 111, 113, 312
 Classical statistical mechanics, 114, 130, 160
 Clip, 31, 257–266, 268–272, 278–281, 286,
 287, 290, 295, 301–304, 306, 307,
 309–319, 322–324, 326, 352, 357–365,
 369–372
 action, 261–263, 265, 266, 268, 269, 272,
 278, 286, 295, 301–304, 306, 307, 315,
 317, 326, 359, 361
 composition of, 258, 265, 281, 290, 301,
 304, 316, 357, 362–363
 fictitious, 258, 358

- Clip (*cont.*)
 percept, 261, 262, 265, 266, 268, 269, 272, 295, 301, 302, 306, 307, 313, 315, 317, 358, 359, 361
 semantic of, 259, 307, 318, 319, 322, 360
 Clockwork universe, 158–161
 Code, 185, 188, 200–205, 240, 282–287, 290, 299, 302
 deformation, 283, 286, 287
 planar, 283
 quantum (*see* Quantum code)
 Shor (*see* Shor code)
 stabilizer, 202–205, 285
 surface, 205, 283–287, 290
 toric, 283
 Coercion, 56, 77
 Coexistence, 49–51
 Cogito argument, 227
 Coherence, 83, 142–145, 148, 193, 197, 220, 224, 229
 Collapse, 36, 39, 112, 123, 126, 149, 168, 204, 221
 Collapse of the wave function, 112, 123, 126
 Collapse postulate, *see* Collapse of the wave function
 Communication, 16, 57, 78, 134, 141, 143, 157, 177, 188, 190, 196, 200, 281, 287–291, 309, 319, 320
 Complementarity, 93, 100, 108, 149–153, 224, 340
 Complete ignorance, 169
 Compulsion, 77
 Computation, 5, 9, 24, 141, 157, 180, 184, 185, 188, 192–194, 236, 281, 282, 285, 289–297, 302, 309, 335, 358
 classical, 184, 285
 measurement-based, 141, 180, 184, 281, 289–297
 quantum (*see* Quantum computation)
 Concept, 4, 5, 15–17, 51, 61, 64, 65, 69, 70, 82, 91, 95, 186, 228, 234, 235, 239, 258, 278, 281, 304, 307, 317, 327, 345, 354, 361, 368–371
 abstract, 258, 281, 304, 307, 368–370
 formation, 281, 317
 Conceptual issue, 31, 92
 Consciousness, 19, 33, 127, 336
 Consequent argument, 63, 121
 Constituent, 2, 9, 21, 30, 33, 162, 220, 238, 336
 Constraining, 34, 77
 Control, 19, 22, 53, 56–59, 61, 65, 67, 68, 74, 77, 79, 80, 82, 116, 119, 136, 138, 139, 143, 144, 148, 157, 163, 168, 171, 174–176, 180, 184, 187, 191, 193, 197, 204, 207–210, 240, 245, 248, 273, 275, 281–283, 287, 290, 298, 315, 347, 377
 adaptive, 80, 210, 240, 281
 diachronic, 57–59
 diachronic direct, 57
 diachronic indirect, 57–59
 foreign synchronic, 56–57
 mechanical, 56
 synchronic remote, 56, 57
 top-down (*see* Top-down control)
 Copenhagen interpretation of QM, 127, 232, 233
 Coping, 11, 20, 334, 337–338, 342–344, 347–349, 351, 353–355, 357–366, 368, 370, 371
 circumspect, 11, 20, 334, 343–344, 347, 348, 351, 353–355, 357–363, 365
 skillful, 342, 351, 354, 357, 368
 Correlation, 7, 49, 50, 118, 145, 153, 156, 157, 164, 171, 173, 174, 177, 181, 185, 188, 191, 235, 288
 Counterfactual scenarios, 80
 Counterfactual thinking, 49, 50, 80, 236
 Cryptographic keys, 189, 190
 Cryptographic protection, 188
 Curie temperature, 4
- D**
 Dasein, 11, 12, 50, 52, 333–338, 340–350, 353–355, 357, 363–372
 Data analysis, 28, 79, 119, 235
 Decision, 12, 20, 60, 80, 220, 221, 230, 247, 255, 257, 259, 268–270, 274, 275, 279, 280, 306, 308, 311, 314, 316, 319, 321–323, 325, 327, 356
 finding, 12, 220, 221, 247, 257, 259, 268–270, 279, 306, 308, 311, 314, 316, 321, 323, 327
 making, 257, 274, 275, 322
 process, 12, 80, 257, 259, 261, 280, 322, 325, 327, 356
 short-time, 322
 situation, 20
 Deep neural network, 306, 318, 319, 321, 325
 Defeater, 21, 54–69, 74, 327
 of agency attribution, 21, 48, 54–65, 68–70, 74
 completely general, 61–65
 general, 61–65, 70
 generic, 66–68
 practical, 61, 62, 65, 67
 psychological, 61

- specific, 55–61
 - theoretical, 61, 62, 65
 - Delayed choice, 204
 - Deliberation, 10, 255, 258–265, 268–271, 279, 280, 307–319, 321–323, 358, 360, 362, 364, 365, 371
 - dynamics of, 10, 259, 262–263, 265, 312, 362
 - history of, 261
 - path, 259, 264, 265, 269–271, 307, 318, 319, 322, 360
 - process of, 258, 260, 261, 263–265, 269, 279, 310, 314–316
 - Dense coding, 174–177, 193, 297
 - Detection, 16, 119, 120, 127, 128, 131, 139, 143, 144, 146, 147, 244, 299, 310, 315
 - Determinism, 9, 21, 25, 32, 61–63, 72, 155, 206
 - super-(*see* Superdeterminism)
 - Deterministic cellular automaton, 62
 - Deterministic dynamics, 21, 160
 - Deterministic itch, 110, 153–187
 - Dirac delta distribution, 124
 - Dirac notation, 105
 - Dirac's bra-ket formalism, 97, 103
 - Directed acyclic graph (DAG), 314, 316
 - Discovery, 92, 222, 235, 238, 239, 281, 287–291
 - Dispersion, 135, 152
 - Double-slit experiment, 119, 140, 143, 145
 - Dutch book argument, 230
 - Dutch book coherence, 229
 - Dynamic equilibrium, 21, 39, 71
- E**
- Eigenbra, 116, 182
 - Eigenket, 116, 131, 132, 182
 - Eigenspace, 106, 112, 121, 122, 202
 - Eigenstate, 106, 111, 112, 116, 122, 124, 128, 129, 135, 150, 151, 153, 175, 177, 179, 181, 182, 192, 201, 245, 291–295, 315
 - left, 182
 - right, 182
 - Eigenvalue, 106, 112, 115, 116, 121–124, 132, 134, 135, 175, 177, 179–182, 185, 198, 201–203, 234, 283–285, 292–295, 297
 - Eigenvector, 105–106, 108, 112, 115, 121, 122, 132
 - Electrodynamics, 92–94
 - Electron, 23, 30, 37, 39, 91–95, 128, 130, 132, 136, 140, 141, 154, 239
 - Elementary particles, 2, 24, 25, 30, 103, 183, 187, 220
 - Eliminativism, 26
 - Embodiment, 19, 71–72, 248, 256, 326, 327
 - notion of, 71
 - Emergence, 9, 31–35, 41, 68–70, 72, 162, 205, 207–210, 309, 345, 346, 349, 371, 377
 - definition of, 32
 - epistemic, 32–34, 208
 - ontological, 9, 33–35, 41, 72, 162, 205, 207–210, 309, 345, 346, 377
 - Emoticons, 263, 268, 269
 - Empirical adequacy, 241
 - Empirical equivalence, 114–116, 127
 - Empirical issue, 31, 52
 - Empirical predictions, *see* Predictions, empirical
 - Empirical question, 155
 - Entangled pairs, 168
 - Entanglement, 32, 143, 145, 156, 164, 174, 175, 184–186, 190, 193, 220, 235, 260, 288–290, 297–299, 301–304
 - Entanglement purification, 289
 - Entity, 8, 19, 21, 30, 31, 34, 35, 41, 47, 49, 58, 70, 71, 73, 78, 191, 208, 220, 224, 256, 313, 326, 327, 338, 341, 358, 365, 367
 - abstract, 19, 26, 46, 49
 - artificially created (*see* Artefact)
 - biological, 34, 68, 78, 186
 - concrete, 34, 46
 - emergent, 72, 208, 232
 - equipment-, 341
 - higher-level, 8, 21, 31, 34, 41, 377
 - physical, 27, 191, 220, 377
 - social, 78, 79
 - Entscheidungsproblem, 335
 - Environment, 5, 10, 11, 16, 18–20, 35, 38, 40, 43, 47, 57, 60, 69, 71–75, 80, 97, 120, 143, 144, 180, 184, 185, 191, 193–199, 204, 205, 210, 220, 222, 223, 238, 249, 256–261, 263, 265, 267, 271–278, 280–288, 291, 292, 297, 300, 301, 305–309, 314, 315, 319, 324–327, 337, 338, 340–343, 347–349, 354–360, 362, 365, 368–372, 377
 - Episodic and compositional memory (ECM), 10, 257–259, 261, 264, 265, 268–273, 278–280, 286, 293, 295, 296, 302–304, 306–319, 321–327, 340, 357–359, 361, 362, 365, 368–372
 - multi-layered, 306–308
 - three-layered, 307, 324, 370
 - Epistemic bookkeeping, 169
 - Epistemological fundamentalism, 227
 - Epistemology, 8, 55, 64, 81

- Equipment, 119, 198, 333, 337–344, 347, 348, 351–355, 357–361, 363, 368, 370
- Equipment, indeterminate, 353–355
- Error, 9, 26, 129, 165, 180, 184, 185, 191–205, 240, 249, 273, 281–288, 291, 292, 302
- logical, 201, 285, 287
- phase-flip, 202
- space, 203, 204
- syndrome, 201, 203, 204, 283–285
- theory, 26
- Event, 1, 6, 15, 16, 20, 47, 60, 83, 92, 93, 119, 120, 159, 168, 173, 180, 206, 210, 225, 299, 305, 309, 376, 378
- Exclusion principle, 95
- Existenciales, 11, 336, 342, 363, 366
- Expectation values, 111, 114–116, 123, 125, 136, 137, 142, 152, 182, 183, 276
- Experience, 1, 11, 74, 75, 80, 186, 221–223, 225, 227, 228, 230, 231, 233, 236, 240, 242, 258, 259, 268, 269, 280, 299, 305, 309, 310, 312, 320, 324, 326, 338, 350–352, 358, 359, 361, 365, 371
- Experiment, 6, 7, 29, 38, 40, 79, 81–84, 93–103, 107, 109, 112, 117–123, 126–129, 131–149, 152–157, 163, 164, 166–169, 171, 185, 186, 188–190, 193, 195, 197, 199, 205–207, 220–223, 226, 227, 232–235, 239, 240, 243, 245, 246, 273, 281, 286, 287, 291–293, 295–304, 306, 307, 313, 317, 324–326, 348, 363, 368–370, 378
- delayed-choice, 142, 143
- double-slit (*see* Double-slit experiment)
- formal description of, 120, 131, 133, 222
- photonic, 281, 297–304, 306
- quantum (*see* Quantum experiment)
- Stern-Gerlach (*see* Stern-Gerlach experiment)
- table-top, 118
- Wigner's thought (*see* Wigner's thought experiment)
- Experimental intervention, 7, 119
- Experimental method, *see* Method, experimental
- Experimental practices, 79, 82, 94
- Experimental science, *see* Science, experimental
- Explainability, 255, 317–319, 321–324
- Explainability, of quantum PS, 322–324
- Explanation, 1, 28, 60, 94, 97, 99, 125, 154–156, 206, 269, 318–321, 323–325, 336, 350
- of action, 336
- of behaviour, 60, 319–321, 324
- long term, 319
- short term, 319
- Explicating, 4, 16–22, 42, 349
- Explication, 8, 16, 21, 22, 42, 43, 50, 71–79, 158, 207, 248, 256, 327, 335, 336, 344, 377
- Explication of agency, *see* Agency, explication of
- Extrinsic, 73
- F**
- Fallibilism, 228
- Feedback, 53, 54, 80, 249, 256, 260, 263, 265, 313–315, 358, 365
- Feynman diagram, 183
- Flexibility, 39, 57, 74, 80, 241, 248, 306, 343, 346, 348, 351, 356, 371
- Forgetting, 245, 264, 265, 267, 278
- Foundations, 9, 23–24, 26, 35, 38, 92, 102, 116, 117, 154, 219, 223, 227
- Frankfurt cases, 77
- Free choice, 84, 171, 207
- Freedom, 18, 20, 75–77, 95, 103, 104, 109, 117, 121, 132, 142–144, 191–195, 204, 207, 235, 238, 239, 258, 260, 265, 270, 291, 299, 371
- Free particle, 113, 123, 124, 169
- Free will, 21, 28, 56, 59, 61–65, 76, 207
- Free will scepticism, 59, 64
- Frege's Basic Law V, 26
- Function, 25, 38, 39, 45, 46, 58, 67, 73, 109, 112–114, 117, 121, 123, 124, 126, 143, 155, 188, 208, 232, 234, 237, 238, 261, 263, 272, 275–280, 294, 299, 305, 306, 316, 317, 319, 325, 327, 356–358, 360, 371, 372
- action-value, 276–278
- conditional probability, 261, 275
- reward, 263, 275, 280, 356
- state-value, 276
- value, 276–280
- wave (*see* Wave function)
- G**
- Gadget, 189, 301, 303, 304, 306, 369
- Gadget formation, 369
- Gaussian distribution, 124
- General relativity, 83

Generator, 179, 189–190, 202, 203, 284, 285, 312
 stabilizer, 179, 202, 203
 true random number (*see* True random number generator)
 Gettier cases, 55
 Glow status, 264, 272, 274, 279, 302
 Grid world, 271–274, 287, 302, 303, 356

H

Hamiltonian learning, 10, 237, 238, 245
 Hamilton operator, 111, 116, 291
 Happening, 2, 6, 15, 16, 22, 28, 29, 34, 40, 47, 49, 51, 53, 61–63, 130, 149, 157, 168, 206, 207, 233, 246
 causal (*see* Causal happening)
 irreversible, 149
 natural (*see* Natural happening)
 worldly, 233, 246
 Heisenberg cut, 120
 Heisenberg equation of motion, 113, 115, 136
 Hermitian adjoint, 105, 107
 Hidden variables, 110, 154–157, 163, 171, 173, 224, 289, 307
 Higher animal, 60, 78
 Higher-level category, 21
 Higher-level entity, *see* Entity, higher-level
 Higher-level structure, 63, 346
 Higher-order organisation, 9, 62, 72
 Higher-order process, 62
 Hilbert space, 103–109, 112, 115, 120, 123, 131, 132, 134, 155–157, 164, 165, 175, 177, 179, 183, 196, 198, 200, 203, 220, 224, 282, 298, 313
 Historicity, 71, 73, 81, 248, 256, 327, 343, 378
 History, 1, 6, 11, 19, 21, 24, 35, 36, 39, 40, 43, 49, 71, 73–75, 80, 95, 102, 150, 155, 159, 186, 205, 222, 227, 258, 260, 261, 275, 276, 303, 309, 310, 319, 323, 326, 327, 335, 340, 347–349, 358, 361, 362, 364, 368, 372, 376, 377
 personal (*see* Personal history)
 of the universe, 36
 Holism, 31–32, 34, 208
 Humanities, 23, 24
 Humean mosaic, 161

I

Idealisation, 33, 241, 358
 Identity of indiscernibles, 186

Ignorance, 69–70, 150, 169, 170
 Immune system, 40–42
 Impact, causal, *see* Causal impact
 Imperative, 47
 Imperfective paradox, 47, 48
 Indeterminism, 9, 20, 34, 38, 41, 62–64, 68, 72, 91, 163, 171, 184, 187, 205–207, 210, 220–222, 248, 256, 270, 308, 309, 327, 343–350, 360, 377
 base-level, 205, 207, 346
 fundamental, 91, 206, 207, 345, 346
 lower-level, 34, 38, 41, 63, 72
 Indeterministic basic dynamic, 10
 Individual, 11, 18, 19, 21–23, 29–36, 38, 41, 42, 58, 59, 68, 70–73, 75, 78, 98, 109, 111, 113, 119, 120, 128, 141, 143, 145, 153, 154, 161, 163–166, 168, 180, 183, 184, 191, 201, 209, 221, 228, 229, 245, 256, 258, 279, 284, 285, 301, 310, 326, 327, 340–343, 351, 354, 355, 358, 361, 364, 377, 378
 concrete material, 71
 development, 229
 ontologically emergent (*see* Ontologically emergent individual)
 traits, 18, 73
 wants, 58, 59, 73
 Individuality, 19, 73, 75, 327, 343, 378
 Influence, 15, 20, 29, 31, 75, 82, 94, 140, 154, 159, 168, 170, 171, 173, 185, 192, 209, 210, 248, 256, 258, 263, 265, 280, 334, 335, 363, 364
 Information, 9, 11, 22, 30, 96, 100, 101, 107, 109, 110, 112–114, 117, 118, 120, 122, 125, 134, 139–141, 143–145, 147–149, 153, 154, 158, 160–162, 165, 169, 170, 174–186, 188–191, 193, 196, 197, 199–201, 203–205, 235, 236, 242, 255, 258, 259, 263, 281–305, 309, 314, 320, 322, 323, 356, 365, 377
 about the system, 100, 109, 112–114, 117, 125, 149, 162
 full, 100, 117, 161, 162, 193, 322
 less-than-perfect (*see* Less-than-perfect information)
 maximal amount of (*see* Pure state)
 which-way (*see* Which-way information)
 Inner mechanism, 59–61
 Inner workings, 57, 59–62, 65, 68, 70
 Instrumentalism, 226, 232
 Intentionality, 19, 20, 39, 51, 72, 336
 Intentional stance, 42, 50, 51
 Intentional states, 20, 43

Interactions, 2, 5, 9, 10, 19, 20, 24, 34, 37, 38, 49, 52–54, 57, 59, 66, 72, 75, 78, 79, 83, 98, 102, 110, 112, 113, 116, 117, 119, 120, 126–131, 137, 139, 140, 143, 144, 148, 149, 171, 173, 177, 183–185, 192–199, 204, 205, 229, 233, 234, 238, 240, 245, 246, 258, 261, 274–277, 280, 282, 286, 287, 291, 292, 297, 303, 306, 309, 312–314, 336, 342, 358, 369, 372, 377, 378

with nature (*see* Nature, interactions with)
physical, 117, 177

Intersubjectivity, 78, 227, 228, 234, 326

Intervene, 7, 57, 82, 207

Invariance, 129

Involvement, 45, 46, 129, 337–338, 340, 341, 351, 353, 354, 361

Irreversibility, 96, 120, 127, 129, 139–140, 145, 146, 196, 203, 204, 221, 242, 309

ontological understanding of, 221

Irreversible process, 130, 139, 147, 191, 209, 221

Ising model, 4, 5, 162

K

Ket-bra, 103

L

Ladder of causation, 49, 80, 236

Language, 3, 17, 20, 21, 43, 45–51, 54, 92, 227, 311, 334, 347, 350, 371

formal, 3

natural, 3

Larmor frequency, 293

Law, of freely falling bodies, 6

Laws, 2, 4, 15, 28, 29, 34, 53, 61, 63, 79, 92, 129, 174, 207, 210, 376

of biochemistry, 2

indeterministic, 34

of nature, 4, 28, 29, 61, 63

Newtonian, 2

of physics, 2, 129

Learning, 5, 6, 8, 10–12, 18, 19, 28, 53, 54, 57, 67, 68, 71, 73–75, 78, 80, 141, 174, 185, 205, 210, 220–223, 228–230, 234–241, 244–249, 255, 257–261, 263–281, 287, 290, 291, 295–297, 302–310, 313, 314, 316–319, 321–327, 334, 335, 340, 342, 343, 346, 348–363, 365, 369–372, 377, 378

agent (*see* Agent, learning)

artificial (*see* Artificial learning)

Bayesian (*see* Bayesian learning)

cycle, 261

deep-Q (*see* Reinforcement learning, value-based)

external perspective, 261

history, 6, 19, 75, 222, 310, 319, 323, 326, 327, 358, 361, 362

how-to, 351–353, 357, 359, 360, 362, 363

long-time, 372

machine (*see* Machine learning)

reinforcement (*see* Reinforcement learning)

through rewards, 318

what-for, 352–353, 357, 362, 363

Less-than-perfect information, 107, 109

Levels

high-, 8, 21, 23, 28–32, 34–36, 38, 39, 42, 43, 50–52, 62–64, 72, 161, 162, 208, 270, 319, 324, 345–347, 349, 360, 368, 377

low-, 15, 21, 25, 31–34, 38, 41, 60, 62, 63, 67, 72, 161, 209, 210, 258, 269, 281, 320, 333, 344–346, 365, 366

microphysical, 30, 31

ontological (*see* Ontological levels)

of organisation, 33, 34, 36, 41, 42, 62, 345, 346

of reality, 21, 25

Life, 4, 8, 16, 19, 22, 23, 27, 30, 32, 36, 40–44, 49, 54, 59, 60, 62, 64, 66, 68, 70, 77, 78, 161, 162, 208, 247, 285, 347, 353, 371

artificial, 43, 66, 78

Lindblad term, 312

Linearity condition, 105

Linear operator, 104, 105, 107, 108, 198

Liouville equation, 114

Living beings, 23, 25, 27, 30, 43, 44, 46, 47, 49, 69, 71, 72, 77, 345

Logic, 20, 24, 26, 81, 178, 201, 202, 204, 206, 221, 282–285, 287, 288, 345

Logical empiricism, 24

Logicism, 26, 29

Loss of coherence, 143–145, 148, 193

Lotka-Volterra model, 4

Lottery, 187, 188, 230

Lower-level material, 31, 62

Lower-level matter, 21, 34

Luck objection, 270

M

Machine learning, 5, 11, 28, 57, 222, 247, 255, 274, 308, 313

Mach-Zender interferometer, 131–148, 314, 316, 322, 326

- Macro-phenomena, 33
- Magnetic field, 4, 94–97, 100, 102, 103, 117, 126, 131, 135–139, 146–148, 151, 165, 167, 175, 176, 193, 210, 232, 239, 243, 245, 291–293, 296, 297, 307, 363
- Magnetic field, homogeneous, 136, 137, 139, 146
- Magnetic field, inhomogeneous, 97, 98, 102, 119, 139, 147
- Markov chains, 311
- Markov decision process (MDP), 261, 263, 274–281, 325, 356
- Markov decision process, partially observable, 261, 325, 356
- Material basis, 345
- Material beings, 15, 16, 19, 29, 30, 347
- Mathematics, 22, 24, 26, 96, 117, 123, 191, 239, 255
- Matter, 2, 9, 15, 16, 18, 20–23, 25, 29, 30, 34–36, 38–42, 44–46, 49, 50, 52, 62, 65, 66, 69–73, 75, 77, 78, 83, 93, 109, 113, 126, 128, 129, 141, 145, 148, 154, 159, 161, 167, 173, 186, 195, 207–210, 220, 226, 229, 230, 234, 236, 248, 258, 287, 327, 335, 336, 345, 346, 350, 377, 378
 - biological, 66
 - lower-level (*see* Lower-level matter)
 - non-biological, 22, 65, 66, 68, 78
 - physical, 248
- Maxwell-Boltzmann distribution, 128, 138, 192, 193
- Measurement, 6, 96, 102, 109–113, 115, 117–123, 126–129, 135, 137, 138, 141, 142, 144, 145, 147–151, 153, 166–168, 170, 171, 175–185, 190, 191, 196, 198, 199, 201, 204, 207, 209, 210, 221, 224, 225, 228, 232, 233, 242, 243, 245, 281, 285, 288–297, 299, 309
 - generalised, 121–123, 199
 - ideal, 122–123, 149, 199
 - problem, 111, 112, 225
- Mechanics, 6–9, 22, 25, 38, 39, 56, 60, 61, 68–70, 79, 83–84, 91–210, 219–248, 255, 260, 291, 296, 305, 309, 311, 315, 317, 322, 323, 346, 356, 377
 - Newtonian (*see* Newtonian mechanics)
- Mechanism, 17, 41, 50, 59–61, 68, 69, 97, 163, 174, 233, 234, 263, 271, 273, 278, 279, 318, 326
- Memory, 10, 12, 18, 19, 21, 41, 60, 71, 73, 74, 162, 180, 184, 200, 204, 205, 221, 247, 248, 255–258, 260–262, 264, 267–269, 271, 272, 278–280, 282, 285, 287, 288, 290, 295, 302, 304, 306, 309, 311, 312, 314, 316, 324, 326, 327, 339, 340, 351, 352, 357, 364, 368, 369, 372, 377, 378
 - of agent (*see* Agent, memory of)
 - episodic and compositional (*see* Episodic and compositional memory (ECM))
 - individual, 71, 73, 377
 - internal, 260
 - quantum (*see* Quantum memory)
- Mental categories, 26
- Method
 - experimental, 6, 7, 38, 79, 81, 82, 94, 96, 98, 100, 118–120, 138, 139, 141, 143, 149, 153, 154, 166, 186, 188, 190, 197, 227, 234, 235, 243, 245, 298, 299, 301, 313, 317, 369
 - philosophical, 1, 3–6, 8, 16, 18, 19, 22–25, 28, 32, 33, 42–44, 55, 60, 65, 77, 79, 81, 91, 102, 117, 161, 173, 189, 226, 228–230, 234, 247, 257, 258, 269, 279, 305, 317, 334, 335, 338, 377
 - statistical, 191, 235
 - variational, 236, 238, 240
- Micro-constituents, 33
- Model
 - of agency (*see* Agency, model of)
 - Bohr's (*see* Bohr's model of the atom)
 - building, 5, 220, 236, 246, 248, 255, 283, 291–297, 307, 317, 324–326, 333, 377
 - of computation, 5, 282, 335
 - of the environment, 10, 249, 275, 277, 283, 287, 288, 325
 - of ferromagnetism, 4, 5, 162
 - formal definition of, 234
 - implicit, 283, 291–297, 317
 - Ising (*see* Ising model)
 - Lotka-Volterra (*see* Lotka-Volterra model)
 - as mediator, 241
 - of Projective Simulation (*see* Projective Simulation, model of)
 - of scientific agency, 6, 10, 234–249, 324, 325, 327, 333
 - as tool, 10, 240–242, 245, 249, 324, 368, 369
- Modelling, 3–6, 8, 10, 12, 15, 17, 18, 22, 42–44, 63, 75, 79–81, 93, 109, 119, 121, 126–129, 162, 180, 220, 222, 233–249, 255, 258, 259, 271, 281, 288, 292, 305, 306, 308, 324–326, 333–335, 344, 346, 349, 352, 358, 369, 372, 377, 378
 - agency (*see* Agency, modelling)
 - perspective, 242, 246, 306, 324–326, 335, 372

- Modelling (*cont.*)
 philosophical, 3–6, 257, 258, 269
 scientific, 3, 10, 234–249, 324, 325, 327, 368, 369
- Monte Carlo techniques, 188, 277
- Moral attribution, *see* Attribution, moral
- Morality, 15, 17, 43, 53, 56, 321, 333
- Moral judgement, 54
- Moral patients, 53
- Moral responsibility, 21, 53, 77
- Moral values, 26
- N**
- Natural
 course, 6, 62, 207
 experiments, 7, 12, 82
 happenings, 2, 15, 22, 28, 29, 61, 62
 language (*see* Language, natural)
 phenomenon, 4, 234, 346, 375
 world, 2, 3, 6, 15, 79, 346
- Nature, 2, 4, 6–10, 15, 21–25, 28–30, 41, 55, 61–63, 68–70, 75, 81–83, 91–94, 113, 117, 119, 130, 140, 154, 158, 168, 173, 174, 186, 191, 206–208, 223, 227, 235, 239–241, 271, 305, 366, 367, 371, 375–378
 discontinuities in, 92
 fundamental theories of, 305
 interactions with, 83
 order of, 2, 15, 22
 representational account of, 7
- Necessity, 3, 23, 44, 70, 75, 78, 163, 344, 345
- Needs, 19, 20, 35, 46, 58, 73, 188, 207
- Newtonian mechanics, 92, 110, 160
- Newton's gravitational theory, 159
- Noise, 51, 60, 123, 129, 143, 165, 180, 184, 189, 191–205, 282, 285, 287–290, 302
 classical, 196–197
 general description of, 197–199
 quantum, 129, 184, 191–205, 282, 285, 287, 290
- Norm, 51, 103, 110, 229
- Normal science, 92
- Normative assessment, 58
- Normative standards, 58
- Nuclear magnetic resonance, 139
- 150–153, 156, 163, 168, 175, 177, 179, 181–183, 185, 192, 196, 197, 199, 201, 202, 204, 206, 224, 228, 245, 283, 284, 290–295, 297
- commuting, 152
 non-commuting, 108
 physical, 106
 spectral decomposition of, 106, 294
 time dependence of, 113, 115, 121
- Observation, 1, 7, 18, 35, 45, 47–49, 79, 81–83, 115, 154, 159, 206, 225, 226, 240, 356
- Observational science, *see* Science, observational
- Observer, 9, 83, 93, 117, 120, 130, 145, 207, 221, 233, 341, 377
- Occam's Razor, 23, 24
- Occurrence, 64, 127, 132, 184, 192, 272, 353, 354
- Ontological category, 23, 25, 27, 29, 42–44, 46, 49
- Ontological emergence, 9, 33–35, 41, 72, 162, 205, 207–210, 309, 345, 346, 377
- Ontological levels, 21–42, 62–64, 71, 72, 91, 127, 207, 208, 346
- Ontologically emergent individuals, 21, 29–42, 70, 72, 73, 221
- Ontological minimalism, 23
- Ontological pluralism, 63
- Ontological pluralist view, 25, 63
- Ontological reality, 21, 36
- Ontology, 8, 19, 21–25, 34, 50, 79, 114, 155, 225, 241, 246, 376
 notions of, 21
 questions of, 22
- Operator, 100, 104–109, 111, 112, 114–116, 120–122, 130–133, 143, 146, 148, 150–152, 165, 167, 169, 178–183, 192, 196, 197, 199, 201–204, 220, 284, 291, 293, 294, 300, 311–317
 commuting, 152, 179
 density, 107, 109, 114, 117, 120–122, 131, 143, 146, 148, 165, 169, 198, 220, 293, 294, 300, 312
 Hamilton (*see* Hamilton operator)
 Hermitian, 105, 106, 108, 111, 121, 132, 152
 identity, 106, 132, 180
 linear (*see* Linear operator)
 non-commuting, 108
 projection, 100, 104, 105, 107, 122, 133, 147, 149, 242
 stabilizer, 179, 183, 204
 unitary, 107, 181, 198, 315
 unitary time evolution, 112, 115
- O**
- Observable, 7, 84, 105–108, 111–116, 119, 121–126, 131, 132, 134, 136–138, 142,

Optimisation, 80, 230, 238, 239, 245, 249
Organic chemistry, 44, 66

P

Participatory realism, 10, 231–234
Path, 2, 37, 41, 44, 74, 80, 137, 140, 142–145, 147, 148, 196, 207, 259, 263–265, 269–274, 302, 303, 307, 310, 312, 318, 319, 322, 339, 360
 deliberation (*see* Deliberation path)
Pauli matrices, 133, 176, 179
Percept, 17, 20, 24, 50, 52, 120, 130, 185, 257–272, 280, 286, 292, 295, 297, 300–302, 305–307, 309, 312, 313, 315, 317–319, 322, 323, 327, 340, 342, 343, 351, 352, 354, 356, 358–362, 364, 365, 371
 clip (*see* Clip, percept)
Personal experience, 225, 227, 230, 242
Personal history, 1
Perspective, 2, 3, 19, 50, 55, 58, 74, 84, 91, 93, 96, 97, 114, 116, 138, 145, 147, 151, 152, 154, 158, 163, 180, 206, 219, 223, 228, 230, 234, 240–242, 245, 246, 258, 261, 280, 287, 298, 299, 305, 306, 324–326, 333–335, 337, 340, 343, 346, 347, 353, 355, 357, 362, 365, 366, 369, 371, 372, 378
 everyday, 2
 personal, 245
Pessimistic meta-induction, 155
Phenomenological analysis, 11, 12, 327, 333, 334, 344, 366
Phenomenological-existential hermeneutics, 343
Phenomenology, 8, 11–12, 234, 235, 255, 324, 333–372, 377
Phenomenology of Dasein, 333, 335
Philosophy, 2, 3, 5, 8, 18, 19, 23–26, 28, 31, 33, 43, 45, 49, 56, 64, 75, 76, 79, 81–83, 158, 161, 162, 228, 231–234, 240, 241, 324, 369, 372, 376
 of mind, 33, 369, 372
 of science, 24, 26, 33, 49, 79, 81–83, 231–233, 240, 241
 of technology, 81
Physical excitation, 262, 313
Physicalism, 24
Physical object, 53
Physical system, 103, 106, 108, 121, 139, 185, 200, 220, 233, 282, 308, 321

Physical world, 1, 118, 231
Physics, 2, 4, 5, 9, 18, 21, 24, 25, 27, 29, 33, 36, 44, 64, 80–83, 92–94, 114, 129, 140, 147, 149, 157, 158, 173, 183, 186, 191, 219–222, 234, 236, 279, 281, 297, 298, 304, 305, 308, 311, 325, 372, 375–378
 atomic, 94
 quantum (*see* Quantum physics)
Planck's constant, 97, 152
Plaquettes, 283, 284, 286
Pluralism, 25, 63
Poisson bracket, 114
Positive operator-valued measure (POVM), 121, 123, 199, 224, 225, 235
Positive operator-valued measure, symmetric informally complete (SIC-POVM), 123, 224, 225, 235
Possibilities
 agential, 344, 346–348
 conceptual, 12, 65, 66
 empirical, 65
 empty logical, 345
 existential, 11, 343–350
 lower-level, 8, 21, 29, 34, 72, 346
 open, 8, 21, 34, 35, 70–72, 80, 171, 327, 343, 345–347, 349, 364, 366, 368
 unactualised, 346
Possible world, 3, 158
Predication, 46
 atemporal, 46
 temporal, 46
Prediction, 48, 81, 92, 110, 115, 117, 125–127, 137, 138, 147, 153, 156, 157, 159, 160, 168, 170, 172, 173, 186, 191, 207, 228, 234, 237, 249, 291, 307, 324–326, 368, 370
 probabilistic, 110, 115, 126, 160, 168, 207, 326
Pre-measurement, 120, 144, 145, 147, 221
Probabilism, 227–230
Probability, 10, 38, 99, 102, 121, 123, 133–135, 137, 139, 142, 143, 151, 152, 156, 168, 169, 172, 178, 192, 193, 195, 196, 199–201, 223–226, 229, 235, 243, 244, 249, 261–263, 266, 268, 270, 273, 275, 276, 285, 289, 293, 295, 296, 310, 315, 325, 326, 356, 361, 377
Born-rule, 224
density, 123, 243, 244
space, 235, 249, 377
transition, 257–259, 262, 266, 275, 296, 308, 314, 317

- Problem of non-locality, 225
- Projection, 5, 100, 104, 105, 107, 110, 112, 122, 132–134, 147–150, 152, 155, 171, 242, 259, 324, 326, 342–344, 348–350, 355, 357, 358, 360, 363–368, 377
- postulate, 122
- Projective Simulation (PS)
- deep, 321
 - formal structure of, 377
 - interpretability of, 318–324
 - model of, 11, 210, 249, 255, 335, 351, 355, 358, 372, 377
 - multi-excitation, 317
 - phenomenological interpretation of, 335, 355–366, 368
 - photonic realisation, 322
 - quantum (Q-PS), 317, 323
 - transparency in, 321–322
- Projective structure, 5, 349
- Property, 4, 5, 100, 102, 104, 117, 122, 149, 155, 203, 224, 234, 239, 275, 319, 340, 341, 358
- emergent (*see* Emergent property)
 - objective, 234
 - structural, 5
- Propositional attitude, *see* Attitude, propositional
- Propositionalism, 227–230
- Propositional judgement, 20, 43
- Purpose, 1, 16, 18, 21, 26, 33, 35, 43, 48, 51, 57–59, 61, 67, 68, 75, 77, 173, 240–242, 245, 246, 248, 305, 337, 340, 352
- foreign, 57–59, 61
 - of its own, 57, 58
- Q**
- QBism, 8–10, 117, 123, 219–249, 255, 326, 377
- QBist update, 10, 224, 230, 235, 245
- Quantum Bayesianism, 223, 226
- Quantum bit, *see* Qubit
- Quantum code, 200–202, 204, 282–287
- in-situ optimisation of, 282–287
- Quantum coding, 200–202
- Quantum coherence, 142–145, 193
- Quantum communication, 134, 141, 143, 157, 177, 190, 281, 287–291
- Quantum computation, 9, 141, 157, 180, 184, 185, 193, 281, 282, 291–297
- one-way, 184, 185, 193, 292
- Quantum computation, measurement-based (MBQC), 141, 180, 184, 281, 291–297
- Quantum computer, 80, 145, 180, 184, 185, 193, 205, 240, 282, 283, 309
- one-way, 185
- Quantum correlations, 118, 156, 157, 164, 174, 177, 185, 191, 288
- Quantum cryptography, 157, 190
- Quantum dynamics, 5, 9, 111–116, 118, 123, 209, 220, 248, 309, 377
- Heisenberg picture of, 9, 111–116, 118, 126–131, 377
 - Schrödinger picture of, 111–116, 118, 126–131
- Quantum entanglement, 32, 143, 145, 164, 184, 185, 191, 289, 290
- Quantum erasure, 142, 144, 145, 147
- possibility of, 144
- Quantum error correction (QEC), 9, 129, 165, 180, 184, 185, 191–205, 240, 273, 281–288, 291, 292
- Quantum experiment, 118–121, 126, 131, 145, 185, 189, 221, 222, 245, 273, 306
- Quantum formalism, 84, 102, 103, 108, 131, 190
- Quantum foundations, 9, 38, 92, 102, 117, 154, 219, 223
- Quantum generalisations, 308
- Quantum holism, 208
- Quantum information, 11, 30, 120, 134, 139, 141, 143, 165, 174–186, 191, 196, 197, 199–201, 203–205, 255, 281, 285, 288, 289, 291, 294, 297, 299, 309, 314, 377
- science, 9, 255, 281–305, 377
- Quantum information, photonic
- implementations of, 140–142
- Quantum information theory, 96, 174–186
- Quantum interferometry, 140, 141
- Quantum Ising interaction, 194
- Quantum measurement, 96, 102, 111, 123, 142, 144, 145, 198, 228
- irreversibility in, 96, 127
- Quantum mechanical tunnelling, 38
- Quantum mechanics (QM), 6–9, 11, 22, 25, 38, 81, 83, 84, 91–210, 219–249, 255, 291, 296, 305, 309, 346, 377
- conceptual foundations of, 102
 - Copenhagen interpretation (*see* Copenhagen interpretation of QM)
 - interactive view of, 233
 - mathematical framework of, 93, 102–116
 - measurement problem of, 112
 - non-relativistic, 38, 39, 155, 163, 173
- Quantum memory, 145, 180, 184, 200, 204, 205, 282, 287, 288, 290, 309, 311

- Quantum number, 95, 141, 239
 - Quantum optics, 10, 119, 136, 197, 234, 279
 - Quantum phenomena, 91, 103, 108, 163, 234
 - Quantum physics, 29, 80, 81, 83, 186, 191, 219–222, 281, 297, 305, 308
 - Quantum projective simulation, 142, 185
 - Quantum random generator, 189–190
 - Quantum randomness, 153–157, 186–205, 210
 - Quantum repeater, 176, 190, 288–290
 - Quantum repetition code, 200–202, 285
 - Quantum state, 84, 109–111, 113, 115–118, 123, 126, 127, 130, 138, 140, 143, 149, 152, 156, 160, 167, 178, 180, 181, 200, 223–225, 228–230, 232, 235, 242, 287, 288, 290, 297, 299, 300, 302, 312, 377
 - attributing a, 109–111
 - entangled, 184, 185
 - status of, 116, 120, 121
 - update of, 10, 116–153, 167, 223, 228, 230, 297
 - Quantum system, 84, 107, 108, 111, 117–121, 123, 134, 143–145, 157, 164, 173, 175, 185, 191, 198, 206, 208, 220, 228, 231, 235, 281, 287, 295, 298, 299
 - composite, 107, 164
 - entangled, 144, 157, 173, 208
 - isolated, 220
 - Quantum technology, 80, 145, 281, 305
 - Quantum teleportation, 30, 119, 175–177, 185, 193, 288, 289, 297
 - Quantum theory, 79, 83, 92, 93, 108, 167, 186, 220
 - Quantum walks, 310–311, 314, 322
 - Qubit, 96, 103, 134–135, 139, 143, 175–185, 193–204, 282–297, 299, 300
 - logical, 178, 200–204, 282, 284, 285
 - single, 139, 175–177, 179, 180, 184, 193, 194, 196, 198, 202, 291–297
- R**
- Radiation, 36, 39, 92, 94, 144
 - Raleigh-Ritz variational principle, 236
 - Random bit, 138, 190, 191
 - Random movie player, 255–258
 - Randomness, 150, 153, 154, 176, 180, 184, 186–191, 193, 205, 206, 210, 256, 309, 377
 - base-level, 153, 189
 - device-independent, 176, 180, 184, 190–191, 193, 281, 297
 - physical, 188–190
 - Randomness, pseudo-, 188
 - React, 16, 36, 37, 40, 41, 49, 52, 64, 68, 74, 80, 118, 128, 130, 139, 153, 154, 192, 236, 246, 291, 371, 375
 - Realism, 10, 25, 30, 70, 173, 177, 226, 242, 246, 326, 376
 - anti-, 226, 231–233
 - local, 173, 177
 - participatory (*see* Participatory realism)
 - Realist, 51, 55, 70, 79, 91, 208, 232
 - Reduction, 26, 27, 29, 30, 36
 - eliminative, 26, 27, 29, 36
 - non-eliminative, 26, 27, 30, 36
 - Reductionism, 21, 25–29
 - eliminative, 27
 - non-eliminative, 29
 - Reductive account, 25, 26, 44
 - Reference, 43, 61, 63, 65, 72, 150, 197, 230, 231, 233, 245, 255, 270, 278, 333–338, 340, 341, 343, 346, 349, 353, 354, 359–361, 367
 - manifold of, 353, 354
 - notion of, 338
 - totality of, 338
 - Reflection, 24, 70, 79, 263, 268, 269, 278, 371
 - Reinforcement learning, 10, 73, 141, 255, 257–260, 263–265, 272, 274–281, 302, 305–308, 314, 325, 355–358, 365, 370, 371, 377
 - model-based, 281, 325, 356–357
 - model-free, 356
 - standard, 10, 255, 258, 272, 274–281, 305, 306, 314
 - value-based, 306
 - Relativity theory, 83, 163, 186
 - Representation, 80, 83, 133, 150, 151, 163, 180, 183, 197, 241, 257, 258, 261, 266, 268, 278, 280, 301, 317, 339, 348, 361, 364, 372
 - Representational account of nature, *see* Nature, representational account of
 - Representationalism, 372
 - Reproduction, 44, 68
 - Reversible process, 130, 147
 - Reward, 73, 260–265, 267, 269, 271–276, 278–281, 293, 295–297, 302, 304–306, 316–318, 322, 327, 356, 365
- S**
- Sceptical argument, 61, 64, 65
 - Scepticism, 7, 9, 28, 39, 40, 54–65, 67, 91, 190, 206, 207, 227, 232, 375, 376
 - agency (*see* Agency, scepticism)

Scepticism (*cont.*)

free will (*see* Free will scepticism)

methodological, 227

Schmidt-Rank Vector, 299, 302

Schrödinger equation, 111, 112, 115

Schrödinger's cat, 220, 221

Science, 3–9, 11, 12, 17, 22–24, 26, 28, 33,
41, 43, 49, 69, 75, 78–84, 92, 118, 139,
154, 155, 185, 186, 188, 220, 226–228,
231–234, 236, 237, 239–241, 247, 255,
281–305, 309, 324, 334, 335, 352, 363,
369, 372, 375, 377, 378

experimental, 7, 118

formal, 24

modern, 6, 82

natural, 6, 23, 24, 26, 155

normal (*see* Normal science)

observational, 7

philosophy of (*see* Philosophy of science)

social, 23, 24, 155

unified, 23, 24

Scientific rationality, 219, 224

Scientific revolution, 24, 82

Self, 19, 20, 50, 60, 69, 70, 78, 158, 206, 207,
221, 222, 230, 247, 255, 305, 318, 333,
338, 347, 349, 355, 366–368, 376

Self-movement, 19, 20, 60

Self-reference, 333, 355, 366, 368

Settler, 18, 20, 73, 343, 345, 378

Shor code, 202–204, 299

Simulation, 5, 22, 75, 80, 141, 188, 249, 258,
259, 287, 301, 302, 309, 312, 313, 317,
324, 326, 357, 364

Social construction, 51

Social interaction, 49, 75, 78, 229

Solipsism, 226

Special theory of relativity, 83

Spectral decomposition, 106, 112, 122, 294

Spin, 5, 95, 97, 99–103, 109, 110, 113, 114,
117, 131–139, 142, 146–153, 155, 157,
164–169, 171, 176, 182, 185, 192, 195,
197, 199, 200, 210, 238, 239, 242–245,
291, 293, 294, 357, 363

degree of freedom, 103, 109

down, 100, 110

measurement of, 102, 110, 137, 147, 148,
210

up, 100, 110

Spontaneous decay, 198

Stabilizer, 179–181, 183, 202–204, 283–285

Stance, 8, 51, 55, 226

intentional (*see* Intentional stance)

naïve realist, 246

realist, 51, 55

State

basis, 111, 135, 165, 193, 196, 200, 201,
294

of belief (*see* Belief, state of)

Bell singlet (*see* Bell singlet state)

Bell type (*see* Bell type state)

cluster, 184, 185, 197, 291, 292, 297

entangled, 145, 165, 175, 176, 184, 185,
202, 281, 288–290, 297–299,
302–304

function of, 117, 155, 208

graph, 180–186, 202, 289, 297, 299, 300

left eigen- (*see* Eigenstate, left)

non-product, 164

orbital angular momentum (OAM), 298

partial trace of, 165

product, 164–167, 180, 201, 299

pure, 109–112, 116, 120, 121, 143, 146,
165, 166, 197–199, 204, 300, 302

quantum (*see* Quantum state)

reduced, 32, 145, 165–167, 170, 182, 194,
284, 300

right eigen- (*see* Eigenstate, right)

stabilizer, 179–181, 202, 203

time-dependent, 113

update of, 110, 127, 133, 223, 230

Stern-Gerlach device, 98–102, 109, 113, 126,
132, 134–139, 141, 142, 146, 147, 166,
167, 210, 242, 243

Stern-Gerlach experiment, 93, 95–102, 109,
117, 119, 120, 128, 129, 131–149, 152,
153, 155, 156, 164, 199, 291

Stochastic dynamics, 279

Structure, 5, 7–12, 20, 24, 28, 29, 41, 50, 60,
63, 71–73, 80, 81, 83, 84, 94–96, 102,
107, 140, 156, 206, 209, 220, 228, 230,
232, 238–240, 248, 255, 260–262, 269,
278, 280, 304, 306, 313, 319, 322,
324–327, 335, 336, 338–340, 342–344,
346, 347, 349–352, 354–356, 358–360,
363, 365, 367–372, 375, 377, 378

categorical, 336

higher-level (*see* Higher-level structure)

input-output, 72

internal, 220, 230, 378

projective (*see* Projective structure)

referential, 338, 342, 355, 359, 360, 366

reflexive, 375

Subjectivity, 18, 19

Sun, 8, 21, 35–42, 55, 70, 71, 127–130, 142,
155, 159, 162, 197, 377

Superconductivity, 33, 174, 186, 247

Superdeterminism, 207

Superposition, 104, 111, 129, 147, 186, 193,
201, 203, 220, 221, 299, 309–311,
316
 quantum-mechanical, 220
Superstition, 27, 29, 36, 51, 188, 375
Supervenience, 161, 162

T

Technology, 6, 30, 58, 79, 81, 95, 189, 190,
305
Techno-scientific practices, 79
Theory, 7–10, 18, 26, 38, 59, 79, 81–84, 91–93,
96, 102, 108, 113, 117, 118, 127, 129,
134, 139, 155–163, 165, 167, 173–186,
189, 190, 192, 205–207, 220, 222–224,
226, 230, 232, 233, 240, 241, 248, 275,
281, 317, 335, 338, 340, 342, 354
Thrownness, 344, 347–349, 355, 357, 358,
363–366, 377
Tool, 10, 77, 79, 108, 109, 127, 152, 157, 160,
162, 163, 205, 219, 222, 226, 231, 233,
234, 236, 240–242, 245, 247, 249, 289,
297, 314, 317, 324, 340, 341, 368, 369,
377
 practical, 233, 234
 theoretical, 233, 234
Top-down control, 208, 210, 377
Toy, 56, 269, 354–355, 363
Traceability, 319, 321–323
True random number generator (TRNG), 189
Turing machine, 5, 335

U

Uncertainty, 108, 124–126, 149–153, 158, 189,
221, 245
Uncovering, 343, 351
Understanding, 2, 6, 7, 20, 28, 36, 47, 50, 51,
56, 59–62, 65, 67–70, 79, 82, 93, 99,
118, 127, 144, 149, 154, 155, 186, 221,
227, 232, 234, 235, 238–241, 247, 324,
335, 336, 342, 344, 345, 348–350, 364,
368, 370, 372, 375, 377, 378
 mechanistic, 70
 physiological, 70
 scientific, 234
Unificationist project, 25

Update, 10, 110–113, 116–153, 166–171, 173,
180, 190, 198, 210, 223, 224, 228, 230,
236, 242, 243, 277, 278, 306, 316, 321
 Bayesian (*see* Bayesian update)
 of belief (*see* Belief, update of)
 QBist (*see* QBist update)
 of quantum state (*see* Quantum state,
 update of)
 of state (*see* State, update of)

V

Van Inwagen's replay argument, 270
Vector, 98, 103–107, 109, 132, 133, 135–138,
142, 152, 155, 156, 164, 167, 195, 220,
244, 291, 294, 299, 302
 arbitrary, 105
 basis, 104, 105, 135
 Bloch, 136
 eigen- (*see* Eigenvector)
 ket, 103, 104
 normalised, 103, 104, 109
 orthonormal, 104, 106
 Schmidt-Rank (SRV) (*see* Schmidt-Rank
 Vector)
Vector space, 103
 two-dimensional, 103
Vienna circle, 24, 227
Von Neumann equation, 114, 116, 311

W

Want, 19, 20, 25, 29, 35, 46–48, 51–54, 58,
59, 71–74, 91, 109, 129, 154, 157, 158,
189, 190, 205, 208, 246, 248, 269, 288,
327, 353, 362, 377
 individual (*see* Individual want)
Wave function, 39, 112, 123, 124, 126, 232,
234, 237, 238
Wave-particle duality, 93, 140, 143, 322, 323
Which-way detector, 143–145, 204
 entangler system, 144
 information, 143, 144
Wigner's friend paradox, 221, 225
Wigner's thought experiment, 221

Z

Zeeman effect, 93–95, 238, 325