Year : 2020

Typicality as a Way of Reasoning in Physics and Metaphysics

Lazarovici Dustin

UNIL | Université de Lausanne
Faculté des lettres

FACULTÉ DES LETTRES

SECTION DE PHILOSOPHIE

Typicality as a Way of Reasoning in Physics and Metaphysics

THÈSE DE DOCTORAT

présentée à la

Faculté des lettres
de l'Université de Lausanne

pour l'obtention du grade de
Docteur ès lettres

par

Dustin Lazarovici

Directeur de thèse

Prof. Dr. Michaël-Andreas Esfeld

UNIL | Université de Lausanne
Faculté des lettres

IMPRIMATUR

Le Décanat de la Faculté des lettres, sur le rapport d'une commission composée de :

Directeur de thèse :

Monsieur Michaël-Andreas Esfeld	Professeur, Faculté des lettres, UNIL

Membres du jury :

Monsieur Sheldon Goldstein	Professeur, Rutgers University | The State University of New Jersey, Etats Unis
Monsieur Barry Loewer	Professeur, Rutgers University | The State University of New Jersey, Etats Unis

autorise l'impression de la thèse de doctorat de

MONSIEUR DUSTIN LAZAROVIC

intitulée

Typicality as a Way of Reasoning in Physics and Metaphysics

sans se prononcer sur les opinions du candidat / de la candidate.

La Faculté des lettres, conformément à son règlement, ne décerne aucune mention.

Lausanne, le 24 juin 2020

Dave Lüthi
Doyen de la Faculté des lettres

Typicality as a Way of Reasoning in Physics and Metaphysics

Dustin Lazarovici
Université de Lausanne

July 5, 2020

Acknowledgements

First and foremost, I want to thank my supervisor Michael Esfeld at the Université de Lausanne, who has guided me over many years and made my transition from mathematical physics to philosophy possible.

I am honored and grateful that Barry Loewer and Sheldon Goldstein agreed to referee this thesis. Their research and teachings – in publications, lectures, and private discussions – have been a great inspiration, and their feedback and comments have been invaluable for improving this work.

Special thanks go to my teacher, mentor, and friend Detlef Dürr. The ideas developed in this work that I owe to him, directly or indirectly, form a set of measure close to one.

I thank my family and friends who were always there for me and close to my heart even when I seemed distant.

And I am honored and grateful to have had David Z Albert as my academic host during a research stay at Columbia University. It was an invaluable experience, both personally and scientifically.

Among the many teachers and colleagues that have influenced my work, I want to thank the following, in particular, for helpful lessons and discussions (with apologizes to those who I forgot to name): Jeff Barrett, Julian Barbour, Christian Beck, Jean Bricmont, Dirk-André Deckert, Saakshi Dulani, Tiziano Ferrando, Mario Hubert, Martin Kolb, Tim Maudlin, Wayne Myrvold, Andrea Oldofredi, Peter Pickl, Paula Reichert, Christian Sachse, Glenn Shafer, Stefan Teufel, Isaac Wilhelm, Gerhard Winkler (†2014), Nino Zanghì.

I thank anonymous referees for helpful comments on previously submitted papers.
And many students whom I've had the privilege of teaching and who, through their questions, objections, and own ideas, pushed me to sharpen mine.

I gratefully acknowledge funding by the Swiss National Science Foundation (SNSF) Doc.Mobility Fellowship P1LAP1_184150.

Several chapters of this thesis are based on papers that were independently published or submitted for publication. Material from publications with co-authors was used with their consent, and only when I made substantial contributions to writing the original text.

- Lazarovici, D. (forthcoming). Typical Humean Worlds have no Laws. Submitted for publication.
- Lazarovici, D. (forthcoming). Typicality versus Humean probabilities as the Foundation of Statistical Mechanics. Submitted for publication.
- Lazarovici, D. (2020). Position Measurements and the Empirical Status of Particles in Bohmian Mechanics. *Philosophy of Science*. DOI: 10.1086/709412
- Dürr, D., & Lazarovici, D. (2020). *Understanding Quantum Mechanics: The World According to Modern Quantum Foundations.* Springer International Publishing. ISBN 978-3-030-40067-5
- Lazarovici, D., & Reichert, P. (2020). Arrow(s) of Time without a Past Hypothesis. In V. Allori (ed.), *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism and Laws of Nature.* World Scientific.
  Preprint: http://arxiv.org/abs/1809.04646
- Lazarovici, D. (2019). On Boltzmann versus Gibbs and the Equilibrium in Statistical Mechanics. *Philosophy of Science*, 86(4), 785–793. DOI: 10.1086/704983
- Lazarovici, D. (2018). Super-Humeanism: A starving ontology. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics.* DOI: 10.1016/j.shpsb.2018.07.001
- Oldofredi, A., Lazarovici, D., Deckert, D.-A., & Esfeld, M. (2016). From the Universe to Subsystems: Why Quantum Mechanics Appears More Stochastic than Classical Mechanics. *Fluctuation and Noise Letters*, 15(03), 1640002.
  DOI: 10.1142/S0219477516400022
- Lazarovici, D., & Reichert, P. (2015). Typicality, Irreversibility and the Status of Macroscopic Laws. *Erkenntnis*, 80(4), 689–716.

Contents

Chapter 1

Introduction

Consider the following regularities that we observe in our universe:

1) Apples do not spontaneously jump up from the ground onto the tree.

2) Rocks thrown on earth fly along (roughly) parabolic trajectories.

3) The relative frequency of *heads* in a long series of fair coin tosses comes out (approximately) 1/2.

These regularities are all of a different kind. 3) is a statistical pattern. 2) is a mechanical phenomenon. 1) turns out to be an instance of the second law of thermodynamics. All three regularities strike us a law-like; arguably, they are even among the more basic experiences founding our belief in a lawful cosmos. However, it turns out that none of them is nomologically necessary under the fundamental microscopic laws that we take to hold in our universe. In fact, given the huge number of microscopic constituents of macroscopic objects and the chaotic nature of the microscopic dynamics, the fundamental laws put very few constraints on what is physically possible on macroscopic scales.

It is possible that particles in the ground move in such a coordinated way as to push an apple up in the air (we know that because the time-reversed process is common and the microscopic laws are time-reversal invariant). It is possible for a balanced coin to land on *heads* every single time it is tossed. And it is possible, as Albert (2015, p. 1) so vividly points out, that a flying rock is "suddenly ejecting one of its trillions of elementary particulate constituents at enormous speed and careening off in an altogether different direction, or (for that matter) spontaneously disassembling itself into statuettes of the British royal family, or (come to think of it) reciting the Gettysburg Address."

Assuming deterministic laws, a physical event or phenomenon is (nomologically) possible if and only if there exist micro-conditions of the universe that evolve under the dynamics such that the event or phenomenon obtains. Given our limited epistemic access to the micro-state of the universe (or any complex system, for that matter) we thus need some inferential procedure from the fundamental dynamics to the salient regularities, other than finding the exact solution trajectory that describes our universe. In fact, even if we *did* know the exact initial conditions and could predict the entire history of the universe deterministically, it would seem odd if law-like regularities such as the ones stated above turned out to be merely accidental, contingent on the very particular micro-configuration of our universe. That is, even if we were Laplacian demons and could verify that dynamical laws + initial conditions make (let's say) the second law of thermodynamics true in our universe, we should care for some additional fact or principle that makes it counterfactually robust and gives it more nomological authority.

The answer proposed in this thesis is a rather simple one. Regularities like those mentioned above (and many others) do not hold for *all* initial conditions – i.e., in all nomologically possible worlds – but the overwhelming majority of them. They are, as we shall say, *typical*.[1]

Typicality

The basic definition of typicality is the following:

Definition. Let Ω be a reference set and Π a set of predicates on Ω.
A property $P \in \Pi$ is *typical* within Ω if nearly all members of Ω instantiate P.
The property is *atypical* within Ω if $\neg P$ is typical, i.e., if nearly none of the members of Ω instantiate P.

For example: The property of being irrational is typical within the set of real numbers. Being a rational number is atypical.

Often, one will hear statements like "a typical real number is irrational." We will adopt this convenient way of speaking from time to time, but be aware that it is a slight abuse of language. No element of a reference class Ω can be typical or atypical per se; it can only be typical or atypical with respect to a certain feature or property (cf. Maudlin (2020)). For instance, the real number $\sqrt{2}$ instantiates the typical property of irrationality but the atypical property of being algebraic.

When applied to a reference class of possible worlds, typicality figures in a way of reasoning about contingency. If a fact about the world is contingent, it means that it could have been different. But not all contingent facts are equally surprising or counterfactually robust or deserving of an explanation. Some facts stand out in that they make our world very special. Some facts could have been different, but only if God – metaphorically speaking – had meticulously arranged things in the world to make it so. Recently, several papers have explored how typicality facts can ground explanations, predictions, and rational belief, both in everyday life and in the context

[1] Other recent publications discussing this concept typicality include Goldstein (2001, 2012); Maudlin (2007b); Volchan (2007); Dürr et al. (2017); Wilhelm (2019); Hubert (2019).

of fundamental physics and statistical mechanics. We will expand on this in detail in the course of this thesis.

Going back to the definition of typicality, one may ask: how do we determine if a subset of Ω (the extension of some predicate P) contains “nearly all” or “nearly none” of the elements? Well, if Ω is finite, then by simple counting. If Ω is an infinite or even continuous set, we usually employ some natural measure in the sense of mathematical measure theory:

$$\texttt{Typ}(P) :\Longleftrightarrow \ \mu\left(x \in \Omega : P(x)\right) \approx \mu(\Omega). \tag{1.1}$$

What makes a measure “natural,” what $\approx$ means more precisely, and what other ways there are to distinguish “very large” and “very small” sets are some of the questions that will be addressed in later chapters. For now, the most helpful answer is that we don’t need to worry too much. In general, all reasonable measures will agree on what is typical or atypical.

Tim Maudlin (private communication) provides the following instructive example: *The Sahara desert is nearly all sand.* One may ask back: “Nearly all by what measure? In terms of area, or volume, or metric tons, ...?” But this means that one didn’t really get the point. The answer is: All of them. Or any of them. By any reasonable standard, the Sahara is nearly all sand. The typical is so overwhelming in number that it leaves little room for ambiguity.

Remark. Following the suggestion of an anonymous referee, I am using the locution *nearly all* for “all except for a set of very small measure” in distinction to *almost all* which usually means “all except for a set of measure zero” in the mathematical literature. The statement *almost all real numbers are irrational* would thus be correct (with respect to the Lebesgue measure and, more generally, any non-discrete measure) but this standard of typicality is too strong in most contexts, in particular for statistical mechanics (unless one considers the idealized limit of an infinite particle number).

From our everyday lives, we are familiar with typicality facts referring to an actual ensemble of entities or events: It is typical for ravens to be black. It is typical for lottery tickets to be a loser. It is typical for calender days not to be a leap-day (though today, at the time of writing, happens to be one).

Such examples convey the right intuition but are of subsidiary interest for our further discussion. The applications of typicality that we will focus on for the most part have a decidedly *modal* character. In particular, the relevant typicality statements in physics refer, in general, to what obtains in most nomologically possible worlds, not to what obtains most of the time in the actual universe. For example: In nearly all possible worlds (consistent with the low-entropy initial boundary condition of our universe), entropy increases on relevant time scales.

The concept of typicality is the same in both cases, but misunderstandings are possible when the relevant reference set is left implicit. For instance, the statement “The 11:45 train from Lausanne to Geneva is typically on time” is arguably true when referring to what happens on most days, but false when referring to what happens (on a particular day) in most possible worlds.[2]

Typicality, in this modal sense, is weaker than *necessity* and stronger than *possibility.* We can, in fact, understand it as a modal operator `Typ` such that

$$\Box p \vdash \mathtt{Typ}(p) \vdash \Diamond p. \tag{1.2}$$

However, $\Box p \vdash p$, i.e., what is necessarily true must be true in our world, while something being typical doesn't logically imply that it actually obtains. Typicality reasoning, as a way of grounding explanations and predictions of actual phenomena, is thus a non-deductive reasoning. It is instead based on the following *rationality principle*[3]:

> Suppose we accept a theory T and we come to believe that our world has the salient property P (expressing, for instance, a physical regularity or phenomenon). If P is typical according to T, there is nothing left to explain. If it turns out that P is atypical according to T, we have to look for additional explanation or, in the last resort, revise or reject our theory.

There is some debate among advocates of typicality whether typicality facts are *predictive*, that is, whether we should endorse a rationality principle like:

> If P is typical according to our theory, we should *expect* P to be instantiated in our world.

I hold that typicality facts are predictive, but the difference between a phenomenon that is predicted by our theory and a phenomenon that is explained when actually observed strikes me as slim, to begin with. The reason for the debate will be addressed in Section 1.2 below.

While the truth of a formal typicality statement depends, strictly speaking, on the measure used to explicate "typical" and "atypical," most of the explanatory work is done by the reference class of possibilities determined by the theory and the laws it postulates – not by any specific measure (a great many choices will make the typicality statement true) and definitely not by any one particular initial condition. Typicality thus figures in a way of reasoning *about the laws*, and it is the laws that ultimately ground (or fail to ground) explanations.

Such typicality explanations are unifying and reductive as scientific explanations are supposed to be: A small set of relatively simple laws and theoretical postulates makes a great number of complex phenomena and regularities typical. However, I won't try to convince the reader that typicality explanations are just a subspecies of a more familiar kind. I rather regard typicality as an elementary way of reasoning, and

[2] Although people who are very fond of both metaphysics and the train service in Switzerland might argue that punctuality is an essential property of Swiss trains.

[3] For this particular formulation, I am drawing a lot from Tim Maudlin (private communication).

in Chapter 10, we will discuss how, in particular, causal explanations turn out to be a form of typicality explanation.

Remark (Conditional typicality versus typicality tout court)**.** Many typicality facts refer not to the reference set of *all* nomologically possible worlds but a subset restricted by pertinent macroscopic boundary conditions. Sometimes, this is quite banal. If we are interested in a physical regularity concerning apples, we consider only possible worlds in which apples exist. But when the boundary conditions are part of the explanans rather then explanandum, it raises questions about their own status. We will have to face this issue in particular when we discuss the need for a special cosmological boundary condition to account for the low-entropy past of our universe (Ch. 11).

One thing we can learn from examples such as *black ravens are typical* is that an instance of a typical property (one particular raven being black) does not warrant any additional explanation that pertains specifically to that instance. We may seek an explanation for the fact that nearly all ravens are black, but once this is explained, there is no interesting story left to tell about why the raven sitting just above my chamber door is black. (A non-black raven, in contrast, would prompt us to explain the deviation from the norm.) In physics, explanations end with the laws. The laws constrain the possibility space of the world. And if a feature of the actual world is typical within this set of possibilities, there is nothing left to explain. To ask further why it is that our world is typical in that particular respect is not only in vain but irrational. One might try to determine the actual micro-history of our universe, but why would that be more explanatory? The actual micro-history is a) unnecessarily detailed and specific to account for the phenomenon and b) contingent on the particular initial conditions of the universe, which are themselves unexplained and most likely unexplainable. So at best, one would establish that P because the history of the world is such that P, which is almost tautological.

What is typical need not *necessarily* happen, and what is atypical is not *impossible*. But assuming our world to be a typical "model" of our theory is basically a necessity of thought. We do it routinely and all the time, if only implicitly. When we are confident not to suffocate because all the air molecules might happen to assemble on the opposite side of the room, we assume typical behavior. When we infer the existence of a tree from our observation of a tree, we neglect the possibility of atypical fluctuations in the electromagnetic field creating an illusion. When we conclude that classical mechanics is falsified by quantum phenomena, we refuse to accept that special initial conditions of a Newtonian universe lead to the observed interference patterns in a double-slit experiment or the violations of Bell's inequality.

On the one hand, it's a condition imposed by us on our theories that they must make the relevant phenomena typical (or at least not atypical). But to the extent that we believe in our theories as (approximately) true descriptions of the world, the fact that we actually observe typical phenomena may fill us with a sense of meaning and gratitude.

1.1 Typicality versus Probability

The previous examples may already suggest that the concept of typicality is related to (and often conflated with) that of *probability*, and large parts of this thesis will be concerned with clarifying the distinction and interrelation between the two. Let me start here with two observations:

1. Contrary to probability, typicality doesn't come in numerical degrees. With respect to a reference class Ω and a set of propositions Π, a property P can be typical, atypical, or neither, but it cannot be *more* or *less* typical than some other property P'. In this sense, typicality is a qualitative rather than a quantitative concept. Even if we use a normalized measure – technically a probability measure – to explicate the locutions "nearly all" or "nearly none," we are not committing to giving meaning to the exact number that this measure assigns to a subset of Ω. The only values that are relevant for a typicality statement are ≈ 1 and ≈ 0.

2. Typicality facts don't presuppose any sort of randomness or indeterminism, nor do they refer to (or depend on) anyone's knowledge or degrees of belief. And as we have seen, typicality facts need not refer to actual frequencies either. In the context of fundamental physics and statistical mechanics, typicality statements express objective facts about the modal structure of the laws, namely which events or phenomena or regularities are instantiated in the vast majority of nomologically possible worlds. These modal facts then come with certain *normative implications* for which theories can be accepted or which phenomena require further explanation.

To summarize in a more catchy motto: When you hear "typical" or "atypical" think, in the first place, of very large and very small sets. When you hear "probable" or "improbable," consult 400 years of debate about what it could mean. That said, the basic – yet at the time revolutionary – idea of predicting what will happen by "counting" possibilities stood at the beginning of probability theory and is also the archetype of typicality reasoning.

It is sometimes criticized that by being content with a typicality explanation or "puzzled" about atypical facts (insisting on additional explanation), we are making an unwarranted inference from typicality to probability, as if, let's say, the fact that a subset of initial conditions is small implies that it is unlikely for one of them to be picked out. Counterexamples to such an inference are readily produced. The bulls-eye makes out a small fraction of a dart board's area, but how likely it is to be hit depends on the skills (and intentions) of the player throwing the darts. Almost all real numbers are irrational, but if we ask a person on the street to name a real number, we would not be surprised if she picked a rational one as those are more familiar to people. When it comes to initial conditions of the universe, there is no one making the pick, no God throwing darts onto the universe's phase space. Thus, some authors argue, we should have no a priori expectations about what our universe is like.

One might point out that in the previous examples, we are already beginning to invoke additional explanations (an agent's dexterity with darts or familiarity with numbers) to account for what seems like a biased distribution of outcomes. And one might further argue that physics is guided by the belief that nature is *not* biased, that if God did play darts, he would put on a blindfold. I wouldn't even reject such intuitions, but it is easy to dismiss them as meaningless or overly romantic.

The more pertinent response is that typicality rather than probability is the right concept for reasoning about the world precisely because it has no connotations of randomness. If a feature of our world is atypical according to a physical theory, it means that our world is – in that particular respect – unlike the vast majority of worlds instantiating the respective laws, the vast majority of models of the theory. It is this fact alone that challenges the theory and creates explanatory pressure. No further inference to probability is made, needed, or even meaningful. And no "a priori expectations" are smuggled in either. The laws determine the set of nomologically possible worlds and thereby typical and atypical properties. And the laws are not a priori, but their empirical (and explanatory) adequacy must be judged in conjunction with typicality. Since special initial conditions could account for virtually anything, a candidate theory can hardly do worse than make the relevant phenomena of our world atypical. Probabilities, on the other hand, are ill-suited for judging a dynamical hypothesis precisely because no initial condition of the universe is "likely" or "unlikely."

In contrast, it can make sense to speak about probabilities associated with throwing a dart or picking a number "at random." But those are physical processes (leaving aside the issue of human consciousness and free will) whose outcome probabilities have to be explained on the basis of the physical laws. In this thesis, we will argue that such explanations are based on typicality. Even though the universe exists only once, and is what it is, the fundamental laws make certain statistical regularities typical. In this sense, the objection that "typicality does not imply probability" has it upside down. Typicality is the more fundamental concept, and probabilistic intuitions are based on typicality reasoning.

The dualistic nature of typicality

Many authors have observed the "dualistic nature" of probability which is often described as epistemic or doxastic on the one hand (referring to incomplete knowledge and/or degrees of belief) and ontic or aleatic on the other (referring to statistical regularities or frequencies). Typicality in physics has a similarly dualistic nature that we could describe as *physical* and *normative.* The first refers to the status of typicality facts as objective facts about the modal structure of the physical laws, the latter to the way of reasoning, i.e., the rationality principles, associated with typicality facts.

Another dualism (so to speak) concerns the *pragmatic* and *nomological* aspects of typicality. On the one hand, there are many practical reasons for resorting to typicality in situations of incomplete knowledge and/or high complexity. We will expound on them in more detail in Ch. 8, but the arguments should be familiar from various introductions to statistical mechanics, which departs from the deterministic description of classical mechanics in favor of an effective description of macroscopic phenomena. However, an important (and rarely appreciated) lesson from statistical mechanics is that even the "deterministic" mechanical phenomena which we describe in classical mechanics, and which form the empirical basis of the theory, are, in fact, typical phenomena. (Whenever we assume that a macro-object can be described as a rigid body or that certain environmental influences can be neglected in a specific situation, we presuppose typical micro-conditions.) This should prevent us from understanding typicality too naively as a concession we make, maybe begrudgingly, to our epistemic and computational limits.

The nomological aspects of typicality concern the issue that stood at the very beginning of our discussion: Typicality clarifies the relation between the fundamental laws and actual phenomena or regularities (falling short of strict necessity), and can ground the nomological status of the latter as *typical regularities.* This issue has nothing to do with epistemic or computational limits. Even the before-mentioned Laplacian demon would learn a new and eminently important fact if he found out that a certain regularity obtains not only for the exact initial conditions of our universe but for nearly all possible ones.

1.2 Atypical versus brute facts

Typicality reasoning applies in a "context" characterized by a reference class Ω and a set Π of relevant predicates. In physics, Ω will generally be determined by the theory and its laws. What determines Π is less clear-cut.

In stating the rationality principle associated with typicality, I spoke, somewhat evasively, of "salient" features of the world, without giving a precise definition of salience. The problem is that, at least as soon as we go to a more fine-grained description, every world is atypical with respect to some properties.

Our universe is certainly atypical with respect to its exact microstate, or the exact number of stars in our galaxy, or the exact position at which the final dart hit the board in this year's world championship. That last week's lottery numbers came out as they did is atypical. And that you, dear reader, are reading this very sentence at this particular time and place is also an atypical event. So why do some atypical features of our world cry out for explanation (or even falsify established theories) while others seem unproblematic and acceptably brute?

The question, what it is that makes a feature of our world salient, a valid target of scientific explanation, strikes me as a very difficult one, and I don't believe that a complete answer can be given in purely physical terms, let alone in terms of abstract measure theory.

Part of the answer is that science is generally concerned with the explanation of robust, usually reproducible phenomena, not individual data points. (Although, across different scientific disciplines, one person's data point may be another one's phenomenon.) For instance, if we shoot electrons through a double-slit onto a photographic screen, the formation of an interference pattern on the screen is a typical phenomenon predicted by quantum mechanics. The exact configuration of impact points in a particular experiment would always be atypical. But it is also not reproducible, not the explanatory target of physicists, and different quantum theories which are considered empirically equivalent disagree on whether the individual impact positions are even in principle determined by physical laws. Of course, I have now merely described a status quo rather than justifying it, and I did not provide a precise characterization of "robust phenomena" as opposed to "data points," so these remarks carry us only so far.

Another observation worth making is that many of our previous examples for atypical events would allow, in principle, for further explanation were we interested in one. The reader might be able to provide reasons for deciding to read this thesis today. And a detailed description of the history of star formation in our galaxy would account for the exact number of stars found today. The drawback of such accounts is that they often trace one atypical event back to other atypical events. But it lies in the nature of causal explanations that they are only shifting the issue to "what caused the cause?" and philosophers rarely complain about them.

Things stand somewhat differently if one wonders why the initial microstate of our universe was exactly X (when there is a continuum of other possibilities all of which are not X). Here, no explanation seems possible and there is something about the question itself that strikes me as deeply irrational. It is the same unease that one might feel about the question of last week's lottery numbers or the count of stars in our galaxy, as well. It had to be *something*, after all, and *any* outcome would be atypical.

Sidney Morgenbesser famously responded to the ontological question: *Why is there something rather than nothing* with: "If there was nothing, you'd be still complaining!" What we want to avoid are such Morgenbesser cases: Why is $F(@) = X$ in our universe? If it were anything else, you'd still be complaining!

So the following addendum to the previously formulated rationality principle seems both necessary and compelling: *The atypicality of a state of affairs creates explanatory pressure only if a typical state of affairs would have been possible.* Otherwise, the fact is acceptably *brute.* More precisely, predicates that do not allow for typicality should not be admitted into Π in the first place.

We need to be more precise, however. The initial microstate of our universe being *not* X is typical; so what makes the predicate $P : \ldots$ *has the initial state* X inadmissible according to the criterion just stated? My point becomes clear if we read X as a variable over possible worlds rather than a rigid designation of one point in phase space. The initial conditions being X is atypical for *any* value of X.

More generally: Let $F : \Omega \to \mathbb{R}^k$, $k \geq 1$ be some function on the fundamental state space of the universe ($\sim$ the set of possible worlds). Then, a predicate of the form

$$P_y(\omega) : F(\omega) = y \tag{1.3}$$

is inadmissible for typicality reasoning if $\texttt{Typ}(\neg P_y)$ for all $y \in \mathbb{R}^k$ (note that an impossible value of F is *a forteori* atypical). In this case, any true proposition of the form $P_y(@)$ would express a brute fact that doesn't require further explanation, at least not based on typicality reasoning.

Consider, in contrast, the predicate $Q(\omega)$: *The second law of thermodynamics holds at* ω. We can restate it (ignoring some subtleties to be later discussed) as the $y = 0$ case in the family of predicates

$$Q_y : \inf \left\{ \frac{\mathrm{d}S}{\mathrm{d}t} \right\} \geq y,$$

so that Q_0 states that the entropy S is non-decreasing. These predicates are admissible because $\texttt{Typ}(Q_0)$ (whereas $\texttt{Typ}(\neg Q_y)$ for $y < 0$). The failure of the second law of thermodynamics would thus not be an acceptably brute fact but severely challenge our theories, arguably to the point of falsifying them.

This analysis also highlights the importance of proper *coarse-graining*. If F is a continuous (or very fine-grained) variable, we are generally interested in predicates that are not quite of the form (1.3) but rather

$$P_y(\omega) : F(\omega) \in (y - \epsilon, y + \epsilon), \tag{1.4}$$

where ϵ must be large enough that a range of typical values exists. For instance, we do not seek to explain why the relative frequency of *heads* in a coin-tossing experiment is *exactly* y but why it is *approximately* $1/2$. (Typically, $\epsilon \propto \frac{1}{\sqrt{N}}$ for a series of N trials.) Similarly, that the number of stars in the Milky Way is exactly N (whatever N may be) is a brute fact, but that the number is somewhere between 10^9 and 10^{14} is, very plausibly, typical for a galaxy of its diameter.

Part I

Probability

Chapter 2

Typicality in Probability Theory

2.1 Expectation Value and Typical Values

> In questions of a practical nature we may be forced to consider events whose probability is more or less close to unity as certain, and events whose probability is small as impossible. Accordingly, one of the most important tasks of probability theory is to identify those events whose probabilities are close to unity or zero.
>
> — Andrey Markov, *Wahrscheinlichkeitsrechnung* (1912, p. 12)[1]

While an important goal of this thesis is to clarify the formal, conceptual, and metaphysical differences between typicality and probability, we shall first discuss the notion of typicality in the context of standard probability theory. This use of typicality is quite appropriate as long as philosophical subtleties and deeper questions about the interpretation of probabilities are left aside. Probability can be a very subtle, controversial, and downright mysterious concept, but the mathematical theory is a rather sober business. In the axiomatic tradition of Kolmogorov, probability theory is nothing but the theory of normalized measures.

Definition. A *probability space* is a triple $(\Omega, \mathcal{A}, \mathbb{P})$ consisting of a set Ω of *elementary events*, a sigma-algebra $\mathcal{A} \subset \mathcal{P}(\Omega)$ of *measurable sets*, and a non-negative, countably additive set function $\mathbb{P}$ on $\mathcal{A}$ – a *measure* – normalized to $\mathbb{P}(\Omega) = 1$. The measure $\mathbb{P}(A) \in [0,1]$ of a set $A \in \mathcal{A}$ is then called "the probability of A."

In our discussion, the elements in Ω will usually describe mere possibilities and $\mathbb{P}$ theoretical probabilities, but a probability space can also be used to describe the actual distribution in a statistical ensemble. Sometimes, hypothetical ensembles are used as a crutch for understanding theoretical probabilities.

Given a measurable function (a so-called "random variable") $X : \Omega \to \mathbb{R}^n$, mapping the elementary events to some numerical value(s), the integral

$$\mathbb{E}(X) = \int_{\Omega} X(\omega) \mathrm{d}\mathbb{P}(\omega) \tag{2.1}$$

[1] Translation from German by D.L.

with respect to the probability measure is called the *expectation value* of X. When referring to an actual statistical distribution, it is called *statistical mean* or *average.*

As the name suggests, the expectation value is generally considered to have a predictive quality even though it need not correspond to the most likely value or even a possible one. As a matter of fact, the expectation value is only a good prediction if significant deviations from it are unlikely, that is, if

$$\mathbb{P}\left(|X - \mathbb{E}(X)| > \epsilon\right) \leq \delta \tag{2.2}$$

for reasonably small values of ϵ and δ. In other words, the expectation value corresponds to a sensible probabilistic prediction only in so far as it provides a good approximation to the *typical values.* We can readily see from (2.2) that there is in general a trade-off between the smallness of ϵ and δ, which, moreover, express inherently vague notions of "significant" deviations and "low" probability. In probability theory and statistical mechanics, this vagueness pops up in different forms and places and becomes a popular point of criticism in philosophical discussions. Therefore, it is important to emphasize from the onset that there is just no way around it. Probabilistic reasoning always involves some degree of vagueness and pragmatism. We'll have to deal with it, no matter what.

A mathematical quantity expressing how much a random variable X fluctuates around the expectation value is the *variance*

$$\mathbb{V}(X) = \mathbb{E}\left((X - \mathbb{E}(X))^2\right). \tag{2.3}$$

The square root of the variance is the *standard deviation*, commonly denoted by σ. The relevance of the variance can be seen from a simple application of the so-called Chebyshev inequality:

$$\mathbb{P}\left(|X - \mathbb{E}(X)| > \epsilon\right) \leq \frac{1}{\epsilon^2}\int |X - \mathbb{E}(X)|^2 \,\mathrm{d}\mathbb{P}(X) = \frac{\mathbb{V}(X)}{\epsilon^2} \tag{2.4}$$

In other words, a small variance ensures a reasonable narrow and thus predictive range of typical values.

2.2 Laws of Large Numbers

The *law of large numbers* – the central result in probability theory – should be understood in exactly this manner. If we consider a family $X_1, \ldots, X_N$ of uncorrelated and identically distributed variables together with their "empirical mean" $m_{\text{emp}} := \frac{1}{N}\sum\limits_{i=1}^{N} X_i$, the variance of the sum is additive (order N) while the pre-factor $\frac{1}{N}$ enters quadratically (as N^{-2}). Hence, (2.4) becomes

$$\mathbb{P}\Big(|m_{\text{emp}} - \mathbb{E}(m_{\text{emp}})| > \epsilon\Big) \leq \frac{\sigma^2}{N\epsilon^2}, \tag{2.5}$$

where we call the expectation $\mathbb{E}(m_{\text{emp}}) = \mathbb{E}\left(\frac{1}{N}\sum_{i=1}^{N} X_i\right)$ the *theoretical* mean.

Again, the trade-off between ϵ (characterizing a small range of values around the theoretical mean) and $\delta(\epsilon, N) \propto \frac{1}{N\epsilon^2}$ (the bound on the probability of larger deviations) is evident. The eponymous "large number" is the ensemble size N. In view of (2.5), it must be large to ensure not that the result is correct, but that it is relevant.

For *Bernoulli variables* $X_i \in \{0, 1\}$ (each outcome either obtains or not) m_{emp} is the *relative frequency.* The law of large numbers then states that *typical relative frequencies* lie in a small range of values around the theoretical mean.

Remark (The $\sqrt{N}$ law)**.** The standard deviation of a sum $\sum_{i=1}^{N} X_i$ of N independent variables (not normalized) is of order $\sqrt{N}$. This characterizes very generally the range of *typical fluctuations* around the mean. Typical *relative* fluctuations are thus of order $\frac{1}{\sqrt{N}}$.

The reason why *statistical mechanics* works so well is that it deals with extremely large N, usually the number of microscopic degrees of freedom in a macroscopic system. In fact, the most important constant in statistical mechanics is not Boltzmann's but Avogadro's constant, $N_A = 6.02214076 \times 10^{23} mol^{-1}$, which is the number of molecules in one mole of a given substance, e.g., in $18g$ of water (H_2O) or $32g$ of oxygen (O_2). Hence, the number of microscopic constituents in a macroscopic system is typically of the order of $N \sim 10^{24}$. This huge number manifests the *separation of scales* between the microscopic and the macroscopic regime, which makes the inherent vagueness emphasized above – the trade-off between ϵ and $\delta(\epsilon, N)$ – unproblematic in practice. Simply put, huge N gives us enough wiggle room to choose both ϵ and δ small enough that typical fluctuations are empirically irrelevant, while atypical large fluctuations are truly negligible on the relevant time scales. (Indeed, time scales are an additional factor to be considered in the trade-off.) On the other hand, when some publications discuss statistical mechanical models with $N = 20$ or $N = 10$ or in some cases even $N = 1$, those have to be regarded with suspicion, to say the least.

Remark (Stronger LLN estimates)**.** Assuming statistical independence, stronger LLN estimates than (2.5) can be obtained from the general form of the Chebyshev inequality

$$\mathbb{P}\left(|Z - \mathbb{E}(Z)| > \epsilon\right) \leq \frac{\mathbb{E}\left[(Z - \mathbb{E}(Z))^m\right]}{\epsilon^m},\ m \in \mathbb{N} \tag{2.6}$$

with $Z = \frac{1}{N}\sum_{i=1}^{N} X_i$ depending on the regularity of the random variables, i.e., in particular, up to which m the *m'th moments* $\mathbb{E}(X^m)$ remain bounded. Exponential bounds, e.g., of the form

$$\mathbb{P}\left(|Z - \mathbb{E}(Z)| > \epsilon\right) \leq e^{-\frac{N\epsilon^2}{const.}} \tag{2.7}$$

for Bernoulli variables are sometimes called *Chernoff bounds.*

Another fundamental result in probability theory is the *central limit theorem* which states that if $(X_i)_{i\geq 1}$ is an independent and identically distributed family of random variables with expectation m and variance σ^2, the distribution of

$$\sqrt{N}\Big(\frac{1}{N}\sum_{i=1}^{N} X_i - m\Big) \tag{2.8}$$

converges to a normal distribution with mean 0 and variance σ^2 for $N \to \infty$. Morally, this means that, for large N, the distribution of $\rho_{\text{emp}} = \frac{1}{N}\sum\limits_{i=1}^{N} X_i$ is approximately a Gaussian centered around m with standard deviation $\frac{\sigma}{\sqrt{N}}$. The relevant observation here is that the distribution becomes more and more peaked with growing sample size N, its weight being concentrated within a few standard deviations of order $\frac{1}{\sqrt{N}}$.

Let's consider the standard coin-toss model, viz. $X_i \in \{0,1\}$ with $\mathbb{P}(0) = \mathbb{P}(1) = \frac{1}{2}$. (Let's say that 0 stands for the outcome *tails* and 1 for *heads*.) The theoretical expectation value for the individual trials is then $\mathbb{E}(X_i) = \frac{1}{2}$, which is rather meaningless as a prediction for a single coin toss. In fact, for a single coin toss, we can say nothing more than that the outcome will be either *heads* or *tails* – which is not saying much at all. Figure 2.1, however, shows the cumulative distribution for the total number of *heads* in a sequence of N coin tosses, for $N = 40$ and $N = 400$, respectively. We can see the Gaussian shape emerging. More importantly, we can see that the distribution is essentially concentrated on outcomes for which the total number of *heads* deviates by at most $\sqrt{N}$ from $N/2$. Correspondingly, a range of typical values for the *relative frequency* of heads and tails is $[\frac{1}{2} \pm \frac{1}{\sqrt{N}}]$.

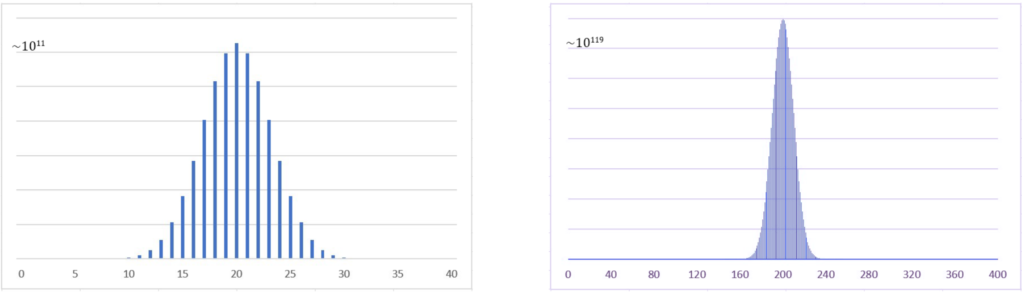

Figure 2.1: Bernoulli distribution with $p = 1/2$ and $N = 40$ (left) $N = 400$ (right).

One of the central claims of this thesis (which may or be may not be controversial) is that the empirical and epistemic import of probabilities in physics comes only from such cases dealing with statistical regularities in reasonably large ensembles. A deeper and more forward-looking observation is that we didn't need to refer to "probability" at all in the coin-toss example. The relevant result is simply that we find a roughly equal number of 0's and 1's in the great majority of possible sequences. This is a genuine typicality fact.

Remark (Kolmogorov's zero-one law)**.** It is a general feature of large families of mutually independent random variables that they partition Ω into very large/small sets corresponding to typical/atypical properties. In a somewhat idealized form (since referring to infinite families), this is manifested in *Kolmogorov's zero-one law*:

Let $(X_n)_{n\geq 1}$ a family of independent random variables on $(\Omega, \mathcal{A}, \mathbb{P})$. An event $A \in \mathcal{A}$ is called *tail event* for the family $(X_n)_{n\geq 1}$ if for all $k \in \mathbb{N}$, its occurence depends only on $(X_n)_{n\geq k}$. Informally speaking, tail events are not sensitive to the values of individual X_n (more precisely, finite subsets of the infinite family of variables) but express "asymptotic properties."

Let $\sigma(X_n : n \geq 1)$ be the smallest sigma-algebra containing all such tail events. Then

$$\mathbb{P}(A) = 0 \text{ or } \mathbb{P}(\mathrm{A}) = 1, \quad \text{for all } \mathrm{A} \in \sigma(\mathrm{X_n} : \mathrm{n} \geq 1). \tag{2.9}$$

A prominent instance of this general result is the *strong law of large numbers* for asymptotic frequencies in the limit of infinitely many trials. Since the strong LLN concerns such (hypothetical) limits, its relevance is more theoretical than practical.

Let us return, for now, to the probabilistic language and emphasize again: What should be taken as the "prediction" of a probabilistic model is not the expectation value but the typical values. When we say "typical values," we mean that the possibility of atypical events – those lying outside the predicated range – can be considered as negligible. An atypical event occurring after all (for instance 900 times *heads* in a series of 1000 tosses) would be the kind of event that compels us to revise or reject our model (e.g., conclude that the coin is biased) rather than shrugging our shoulders and say: "Well, I guess anything is possible." In practice, of course, one may choose to be a more aggressive predictor and report a smaller range of predictions for the price of greater uncertainty. Whether this is a reasonable thing to do depends not only on the rarity or improbability of the neglected outcomes but also on their potential impact.

Note that what we refer to as the range of typical values (specified above by ϵ) reflects an "intrinsic uncertainty" of the probabilistic model that comes, ideally, with rigorous estimates as those provided by the (weak) law of large numbers. It would be misleading to call it a "margin of error" in so far as this suggests something faulty with the input data or the mathematical derivation. In practice, limited measurement accuracy, error propagation in numerical simulations, etc. would rather come in as additional sources of uncertainty (unless, of course, they are precisely what the probabilistic model is supposed to model). Needless to say, nothing is as misleading as predicting some exact value without any estimate of potential errors or typical fluctuations (this omission is usually a good indicator of a "soft science").

In any case, a probabilistic model always yields a *range of predictions* or *prediction interval* (though not necessarily a single connected one). In some unfortunate cases, this range may be so large that the model is not very predictive at all. In particularly nice cases, the range of typical values will be narrow and centered around the expectation value so that the latter becomes a good representative of the predicted range. This applies, in particular, in the regime of the law of large numbers or the central limit theorem (that is, when we can assume a reasonably narrow Gaussian distribution) but it can fail spectacularly for systems with strong correlations or variables with large, or even unbounded, variance. In fact, it has been argued that many catastrophic prediction failures in finance and other socio-economic domains arise from unjustified assumptions of Gaussianity, which lead to underestimating the likelihood of extreme events (Taleb, 2010). For instance, a severe market crash appears to be less improbable than assumed in established risk models, and even if it occurs only once every couple of decades or so, it may wipe out all the cumulative profits that a bank has generated in the meantime. On time scales relevant to a financial institution, it is thus a rare but by no means negligible event.

An analogous observation applies to actual statistical distributions. If we consider a sample consisting of 1000 nurses and Warren Buffett, the *average* income and the *typical* income will differ significantly. (The *median* income would be a better approximation to the latter, but that need not be true in general.) In such cases, it seems evident that (2.1) is not necessarily a good guide-post for practical decisions.

The following is a textbook example[2] of a game with positive expectation value in which the player typically loses money in the long run. It is also known as the *St. Petersburg paradox.* In the limit of $n \to \infty$ rounds, the expectation value is even infinite, while the risk of ruin is 100%. Would you take the bet if you had to commit to a game of (let's say) $n = 1000$ rounds? If we agreed that you should not, we would seem to agree that rational decisions/expectations are based on typical outcomes rather than expectation values.

Example (A game with infinite expectation value in which you will typically go bankrupt)**.** Consider a biased coin for which the probability of *heads* is $p \in (1/3, 1/2)$. You start with a positive capital of X_0 dollars. In each round, you double your capital if the outcome is *heads* and lose half of your capital if the outcome is *tails.* The expectation value for the $(n+1)$'st round is thus

$$\mathbb{E}(X_{n+1}) = p(2X_n) + (1-p)\left(\frac{1}{2}X_n\right) = \left(\frac{1}{2} + \frac{3p}{2}\right)X_n = \left(\frac{1}{2} + \frac{3p}{2}\right)^{n+1} X_0 \xrightarrow{n\to\infty} +\infty.$$

However, the probability of obtaining m times *tails* on n trials is $\binom{n}{m} p^{n-m}(1-p)^m$, resulting in a capital of $2^{(n-m)}\frac{1}{2^m}X_0 = 2^{n-2m}X_0$. The typical values of m for large n are $\frac{m}{n} \approx (1-p) =: \frac{1+\delta}{2}$ and thus $X_n \approx 2^{-\delta n}X_0 \xrightarrow{n\to\infty} 0$. Indeed, a rigorous proof

[2] Figuratively and literally, see, e.g., Georgii (2004, Ex. 5.8); for a recent philosophical discussion in the context of typicality, see Maudlin (2020).

using the strong LLN shows that $\lim_{n\to\infty} X_n = 0$ *almost surely.*

2.3 Interpretations of Probability

We have taken for granted that probabilities are somehow related to (rational) expectations, but omitted the question, what they are actually referring to. This is a famously difficult and controversial topic. I believe that different concepts of probability – when clearly distinguished – can peacefully coexist in separate contexts. This thesis, however, is mostly concerned with probabilities in the natural sciences, in particular physics, which must be somehow connected with natural laws on one side and empirical phenomena on the other. I shall start by addressing two possible meanings of probability that will *not* be the focus of our further discussions.

Subjective Probabilities

Objective and subjective concepts of probability have existed in parallel and opposition since the early days of probability theory (see, e.g., Hacking (1975)). The notion of probability that we are going to discuss in this thesis will fall on the objective side.

Surely, subjective probabilities have their place in everyday reasoning, maybe even in some special sciences and certain areas of philosophy. When I say: “There is a 30% chance that my ex will respond to my text message,” I really mean: “My credence that my ex will respond to my text message is (roughly) 30%.” Such an estimate may be based on objective factors or prior experience but ultimately expresses a personal degree of belief. (Hopefully, at least, I’m not as desperate as to produce a statistically relevant sample of text messages.)

Probabilities in physics do certainly have an epistemic and behavior-guiding function, as well. But first and foremost, physics should seek to explain why *as a matter of fact* certain statistical regularities obtain, be it in a series of coin tosses or a particle scattering experiment. If our theory is able to explain and predict such objective regularities, I begin to see how one could justify the rationality of assigning credences about individual instances – e.g., the outcome of the next coin toss – accordingly. I fail to see, however, how the dispersion of heat or the creation of an interference pattern in a double-slit experiment could depend on anyone’s knowledge or degree of belief. We have already remarked on the “dualistic nature” of probability, and as Myrvold (2016) points out, we really need both the epistemic and the ontic (I’d rather say “factual”) aspects. My skepticism towards epistemic probabilities in physics concerns precisely their ability to bridge the gap to the factual level.

I remember how once, as a young student, I argued: “When I say that the probability of the coin landing on *heads* is 1/2, I do so because I don’t know the exact initial conditions that would allow me to predict the outcome deterministically.” One of my former colleagues, Christian Beck, replied: “Even if you did know the exact micro-conditions, it wouldn’t change the fact that if you throw the coin multiple times, it will land on *heads* and *tails* roughly equally often.” This was an eye-opener for me. Indeed, it is such facts that physics is tasked with explaining and for which a probabilistic language is appropriate even when a deterministic account is available, in principle. Here, I should alert the reader that in the view to be developed in this thesis, probability will appear rather on the side of the explanandum than the explanans – which makes its objective character only more evident.

Of course, concepts like “explanation” and “prediction” have a psychological dimension. We make certain inferences based on our best available theory, from which we conclude that some phenomenon is to be “expected” or at least “unsurprising” if actually observed. It is thus tempting to think that such inferences take epistemic or doxastic states as input, that we deduce knowledge or belief from prior knowledge or belief. This view strikes me as fundamentally misguided, though. Explanations and predictions based on physical theories should be statements about physical facts and laws. The respective inferences will often involve pragmatic considerations, approximations and idealizations, but they should not involve mental states. The psychological dimension comes in at a later point, and lies, strictly speaking, outside the purview of physics. It concerns the *normative implications* of objective physical results, as we will discuss at various points throughout this thesis. Note that even logical deductions have such a normative aspect to them: we *should* accept the conclusion if we accept the premises. But there is neither a natural nor man-made law compelling us to do so.

One of the most pernicious effects of subjectivist interpretations of probability in physics is the idea that the role of probability measures is somehow to “guess” the actual microscopic state of a system (if not the entire universe). The argument goes something like this: If we knew the precise initial conditions of a deterministic system, we could solve the equations of motion exactly (at least in principle). Since we don’t know the precise initial conditions, we put a probability distribution on the pertinent state space (or a subset thereof), which expresses our knowledge/ignorance about the system’s actual microstate. Hence, the more concentrated this probability distribution, the more information we have about the system. However, probabilistic reasoning, at least in statistical mechanics, applies in general to *macroscopic* histories and *macroscopic* regularities. If λ_A is the equidistribution on a phase space region A, and λ_B the equidistribution on a strictly smaller region $B \subset A$ it does not imply at all that the measure λ_B contains “more information” because it localizes the microstate in a smaller subset of phase space. In fact, it may give us *less* relevant information by assigning significant weight to a greater number of macro-states/macro-histories or less weight to the actual one.

Consider, for instance, the situation depicted in Fig. 2.2 where the small set A_1 of initial micro-conditions leads to macro-history 1 while the larger complement $A \setminus A_1$, containing the actual initial state, leads to the macro-history 2. Suppose a demon tells

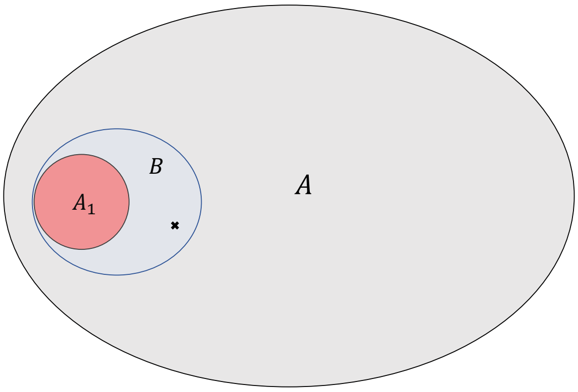

Figure 2.2: Initial conditions in A_1 lead to the macro-history 1, while initial conditions in $A \setminus A_1$ lead to the macro-history 2. The x marks the actual microstate contained in $B \cap A \setminus A_1$.

you that the actual microstate is contained in the set $B \supset A_1$. You have learned a new true fact about the world and significantly narrowed down the set of possible initial conditions. Yet, the demon (being evil) has made you much *less* informed about the macro-history that the system will, in fact, undergo.[3]

That said, I believe that the notion of "information" is even more ambiguous – and more often abused – than the notion of probability, so that explicating the latter in terms of the former is not helpful, in general.

To end this discussion on a more conciliatory note, we note that there are certain parallels between the concept of typicality advocated in this thesis and the view of Myrvold (2012, 2016, 2020) – even though the latter has a more epistemic flavor. To draw these parallels, one could say that typicality facts correspond, for all practical purposes, to those probabilistic predictions on which all reasonable assignments of prior probabilities agree. The latter are what Myrvold identifies as the relevant predictions of statistical mechanics. I believe, however, that at the end of the day, the subjectivist connotations are doing no good, even when they are constrained by (objective) rationality principles or figure only in an intersection of beliefs. The relevant predictions that can be derived in statistical mechanics are those which hold true for nearly all possible microstates. These are objective nomological facts. And what all reasonable measures agree on is the meaning of "nearly all."

Indeterministic Laws

Throughout this thesis, I won't say much about indeterministic theories, or, more precisely, fundamentally stochastic laws that involve irreducible randomness. There

[3] Cf. also Shafer (1985) on the related problem of conditionalizing probability.

are certainly interesting philosophical questions about such probabilities, in particular, the question, what they are actually doing in the world, how they relate to the concrete phenomena. These questions should be taken seriously – so seriously, in fact, that I don't think indeterminism is of much help in understanding probability – but they are of a somewhat different kind than the issues at the center of this thesis.

The one important observation that I do want to make is that even if we consider stochastic theories that assign nomic probabilities to possible events, their empirical and explanatory import is ultimately based on a form of typicality reasoning or Cournot's principle of the negligible event that we will introduce in the next section. Indeed, the necessity of such a principle comes out very clearly in the indeterministic case.

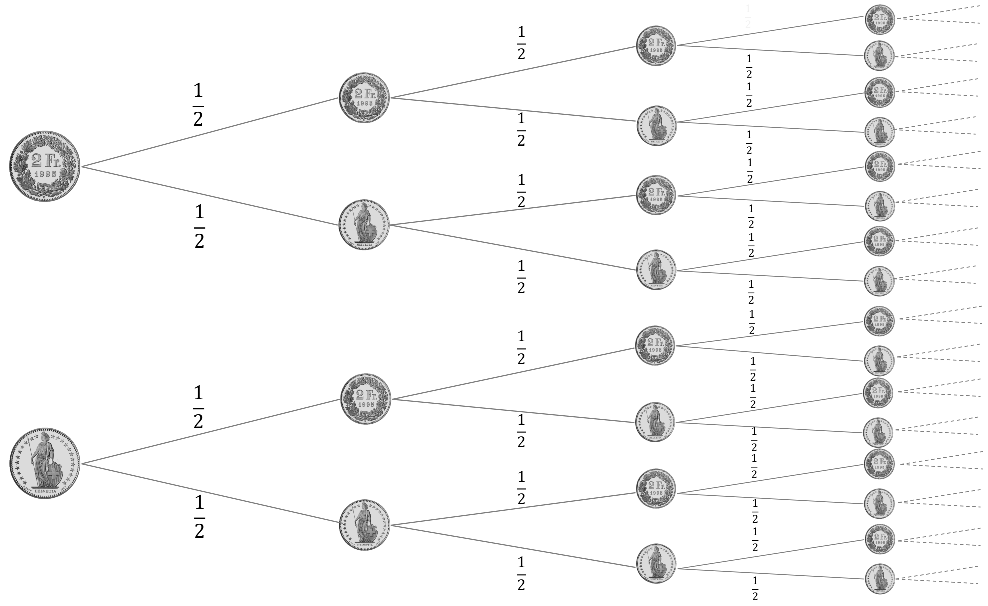

Figure 2.3: Branching structure of possible histories after 4 coin tosses. The weights correspond to the fundamental probabilities assigned by a stochastic law.

Let's consider again the simple example of a series of N coin tosses (the reader may also think of spin measurements on a spin-1/2-particle if she prefers), but conceived as intrinsically random events. That is, for each possible trial, the laws of nature do not determine a unique outcome given the complete initial state but assign only a probability of 1/2 to each of the possible outcomes 0 (*heads* or *spin down*) and 1 (*tails* or *spin up*). The possibilities for the first 4 iterations of the experiment are depicted in Fig. 2.3. We get a branching structure of possible histories with the laws determining the weight or probability of each branch. Each history, i.e., each conceivable sequence of outcomes, is possible and, in fact, equally likely. So in what sense is the law even predictive? Call this the *problem of vast possibilities.*

What we should take a stochastic law to predict is not any *individual* history but the regularities to which it assigns a very high cumulative probability. In the present example, all possible branches have equal weight, but the set of branches in which the relative frequency is approximately 1/2 sums up to a total weight that is very nearly one (for large N). In other words, for the law to have any empirical implications at all, we must understand it as saying that the *typical* frequency of 1/2 will, in fact, be observed. Or, conversely, that atypical histories (such as 0 coming out on every single trial), though not impossible, are not going to occur.

The problem of vast possibilities is only more pronounced when we consider more realistic proposals for indeterministic laws since stochastic micro-dynamics will make virtually any conceivable history of the world *possible*. So again, we should take the regularities that such a law predicts or explains to be those that come out as typical, i.e., those to which it assigns a probability close to 1. If the observed phenomena turn out to have a very low probability, i.e., come out as atypical, it is not strictly speaking a violation of the law. Yet, the rational consequence would be to consider the law as falsified, that is, empirically inadequate.

As we will discuss in detail, the situation is actually pretty much the same for deterministic theories. Even if the initial micro-conditions determine one unique history, these initial conditions are never known exactly. And even if they were, one should question if they would be particularly explanatory (for why was the initial microstate of the universe exactly what it was?). The crucial difference between stochastic and deterministic dynamics is that the latter do not assign probabilities to possible histories. What is *typical* according to deterministic laws is rather what obtains *for nearly all* possible initial conditions. This raises the question, what measure we should use to quantify microstates.[4] It is one of the questions that will be addressed in this thesis.

In the upshot, the central aspects of typicality reasoning that we will develop against a deterministic backdrop apply quite analogously in the indeterministic case, while indeterminism and fundamental randomness are of no help with the main concerns of this thesis but only raise additional problems.

Some readers may object that our focus on determinism is not very naturalistic: Aren't our best physical theories, viz. quantum theories, indeterministic? As a matter of fact, they probably aren't. Standard quantum mechanics involves only one precise dynamical equation, the Schrödinger equation describing the time-evolution of the wave function, which is perfectly deterministic. Randomness comes in only with the measurement process and the infamous collapse postulate, according to which the "measurement" of an "observable" produces one of the possible outcomes (eigenvalues of the associated operator) with probabilities given by the Born rule. But what exactly qualifies as a "measurement"? What are the precise physical conditions for the Schrödinger evolution to be suspended in favor of collapse? And how exactly does the measurement process produce a definite outcome? The standard story is hopelessly vague, as no one pointed out more clearly than John Bell (2004) despite earlier

[4] In fact, this question might also arise in indeterministic theories if the stochastic laws determine only transition probabilities between states but no probability distribution over initial states.

complaints by Einstein and others.

More precisely, standard quantum mechanics (and, in the same vein, quantum field theory) is plagued by the *measurement problem*, that is vividly illustrated by Schrödinger's infamous cat paradox. We will discuss it in more detail in Chap. 12. There are essentially three precise quantum theories (or classes of theories) that solve the measurement problem and ground the quantum formalism in an objective description of nature: Bohmian mechanics, spontaneous collapse theories (such as GRW), and Many Worlds theories. Two of those – Bohmian mechanics and Many Worlds – are fundamentally deterministic; we will study them in more detail in Chs. 8 and 12, respectively. Collapse theories are indeed indeterministic, i.e., involve real irreducible randomness. Somewhat ironically, however, this is also the class of theories whose predictions deviate, in principle, from what are taken to be the predictions of orthodox quantum mechanics. Hence, one could say that if quantum mechanics is indeterministic, it is not exact, and if quantum mechanics is exact, it is not indeterministic.

Chapter 3

Cournot's Principle

We are getting close to the modern concept of typicality if we understand the probability calculus in connection with *Cournot's principle.* Cournot's principle (CP) has been somewhat forgotten in modern times (with Glenn Shafer being one of the few famous probabilistis holding up its banner), but has a long tradition in the philosophy of probability, with some version being endorsed by Kolmogorov, Hadamard, Fréchet, Borel, among others (see Martin (1996); Shafer and Vovk (2006) for excellent historical discussions).
One way to introduce CP is as a remedy to the following dilemma:

> Only probabilistic facts follow from probability theory.
> There are no genuinely probabilistic facts in the world (any possible event either occurs or not).
>
> ---
>
> No facts about the world follow from probability theory. $\therefore$

The second premise could be denied by admitting something like propensities into the physical ontology. But then replace "(physical) facts" by "empirical facts" in the conclusion, and we end up with a similar problem: logically, no empirical facts follow from probabilistic ones. To emphasize, one cannot derive from the probability calculus that an event with probability p will occur (roughly) Np times on N independent trials, only that it will do so with high probability.

Cournot's principle can thus be understood as a sort of "bridge principle," leading from probabilistic results to physical/empirical predictions. An unfortunate historical fact is that the formulation provided by its namesake sounds plainly wrong, at least to modern philosophers of science:

> "A physically impossible event is one whose probability is infinitely small. This remark alone gives substance – an objective and phenomenological value – to the mathematical theory of probability." (Cournot, 1843)

It seems clear from his further writing that Cournot understood "physically impossible" in more of an FAPP (for all practical purposes) sense, referring to a *negligible* possibility (as Borel would later put it) rather than events *forbidden* by natural law. Undoubtedly though, Cournot's terminology has not helped the acceptance of his philosophy.

A more appropriate formulation of the principle can be found in the work of Kolmogorov (1933, Sec. 2.1) : if an event has a very low probability, *"then one can be practically certain that the event will not occur"*. Equivalently (by contraposition): if an event has a very high probability, we should *expect* it to occur. Other authors have cast the principle in more decision-theoretic terms. Roughly: it is rational to act as if very high probability outcomes will obtain.

CP can be applied both to one-shot events (limited to a particular time and place) and complex ones, corresponding to a statistical pattern. In the latter case, it is generally tied to a law of large numbers. We must, of course, be careful not to confuse the relevant referent. It is very likely for a low-probability event to occur *eventually* on repeated trials. Moreover, what counts as "very unlikely" in the sense of the principle can be context-dependent, as we will discuss in more detail below.

Especially in science, when we are dealing with robust phenomena, the most pertinent formulation of CP is in terms of *explanation*:

If a phenomenon has very high probability according to our theory, we should consider it to be conclusively explained by that theory. If we observe a phenomenon of very low probability, we should look for further explanation and possibly (if the phenomenon is significant enough) revise or reject our theory. The reader will have certainly realized that this mirrors the rationality principles that we have set out for typicality.

In the end, I believe that these various formulations of CP are just different aspects of the same rationality principle, which may be cast in terms of expectation, belief, acceptance, explanation, etc. – at least on the level of granularity at which these concepts are roughly intertranslatable.

Example. Atypical events are still *possible* but usually the kind of events that we would not accept as a fluke but that cry out for further explanation and thus challenge our theoretical assumptions. It is possible to roll a *six* 80 times on 100 trials but the rational conclusion is that the die is loaded. My favorite example for the application of CP comes from Martin Scorsese's movie *Casino* (1995). Ace Rothstein, played by the great Robert DeNiro, is a gambling expert hired to manage a Las Vegas casino controlled by the mafia. In a pivotal scene of the movie, he gets into the following exchange with his employee (Don Ward) in charge of overseeing the slot machines. I apologize in advance for reproducing the colorful language.

> Ace Rothstein: *Four reels, sevens across on three $15,000 jackpots. Do you have any idea what the odds are?*
>
> Don Ward: *Shoot, it's gotta be in the millions, maybe more.*

> A.R.: *Three f***in' jackpots in 20 minutes? Why didn't you pull the machines? Why didn't you call me?*
>
> D.W.: *Well, it happened so quick, 3 guys won; I didn't have a chance...*
>
> A.R. [interrupts]: *You didn't see the scam? You didn't see what was going on?*
>
> D.W.: *Well, there's no way to determine that...*
>
> A.R.: *Yes there is! An infallible way, they won!*
>
> D.W.: *Well, it's a casino! People gotta win sometimes.*
>
> A.R. [grows more irritated]: *Ward, you're pissing me off. Now you're insulting my intelligence; what you think I am, a f***in' idiot? You know goddamn well that someone had to get into those machines and set those f***in' reels. The probability of one four-reel machine is a million and a half to one; the probability of three machines in a row; it's in the billions! It cannot happen, would not happen, you f***in' momo! What's the matter with you? Didn't you see you were being set up on the second win?*
>
> D.W.: *I really think you're overreacting...*
>
> A.R.: *Listen, you f***in' yokel, I've had it with you. I've been carrying your ass in this place ever since I got here. Get your ass and get your things and get out of here.*
>
> D.W.: *You're firing me? [...] You might regret this, Mr. Rothstein.*
>
> A.R.: *I'll regret it even more if I keep you on.*
>
> D.R.: *This is not the way to treat people.*
>
> A.R.: *Listen, if you didn't know you were being scammed you're too f***in' dumb to keep this job, if you did know, you were in on it. Either way, YOU'RE OUT!*

3.1 The Lottery Paradox and Rational Belief

One could argue that CP comes essentially for free if probabilities are understood as *credences*. However, as explained before, we are less interested in deriving beliefs from prior beliefs, and more in objective probabilities that derive from our best theories of nature. Such predictions come with certain *normative* implications for which phenomena require (further) explanation and which theories can be accepted. That very high credence means FAPP certainty is basically analytic, while the rationality principle we are interested in derives its normative power, at least in part, from the nomological authority of natural laws. Moreover, epistemic or doxastic theories assign the same status to all degrees of belief, whereas it's characteristic of the view associated with Cournot that statements of "very high" and "very low" probability are privileged with respect to their physical, empirical, and epistemic content. Thus Borel's famous credo: "The principle that an event with very small probability will not happen is the only law of chance." (Borel, 1948)

If one wants to conceive of Cournot's rationality principle in doxastic terms, one could say that it doesn't tie objective probabilities to degrees of belief (like Lewis' *Principle Principle* to be discussed in 5) but "very high probability" to *belief simpliciter*. One problem that then arises is that rational belief is generally assumed to be closed under conjunction

$$Bel(A) \wedge Bel(B) \implies Bel(A \wedge B), \tag{3.1}$$

while typicality statements – or statements of "high probability" – are not. (Unless one applies CP only to events of measure 1.) Clearly, the probability of $A_1 \cap A_2 \cap \ldots \cap A_n$ could get arbitrarily small with increasing n, even if the probability of each A_i is very large.

The account thus faces the infamous *lottery paradox* (Kyburg, 1961): It is very likely for any lottery ticket to be a loser. Hence, I should believe that ticket 1 will lose, and that ticket 2 will lose, ..., and that ticket N will lose. But then, by (3.1), I should also believe "ticket 1 will lose, and ticket 2 will lose,..., and ticket N will lose" – which is certainly false if one of the tickets will be drawn for sure. Such cases in which a conjunction of typical facts ceases to be typical, or even becomes atypical, threaten to make Cournot's principle inconsistent with normal doxastic logic.

A possible reaction is to concede that belief – or at least the relevant notion of "expectation" associated with CP – is not closed under conjunction. Indeed, one may very well read the lottery paradox not as an argument against CP but, on the contrary, against the insistence that rational belief must satisfy (3.1).

If we think in terms of *explanation*, one might also be puzzled at first: Could it be that theory T explains A and T explains B, yet T doesn't explain $A \wedge B$? But yes, it easily could. According to nuclear physics, no further explanation is needed for the fact that plutonium atom 1 hasn't decayed within the past hour, or that plutonium atom 2 hasn't decayed within the past hour, etc. (or, for that matter, that neither atom 1 nor atom 2 have decayed). But no atoms decaying in one $1kg$ of plutonium would cry out for further explanation. We face, of course, a "sorites problem" (what threshold amount of Pu would make it "puzzling" if no decay occurred?), but we established from the very beginning that typicality – and probabilistic reasoning, in general, – is vague.

A more sophisticated response was developed by Hannes Leitgeb (2014, 2017), who tackled the problem of reconciling normal doxastic logic and degrees of belief (satisfying the axioms of probability) with the "Lockean thesis": there exists a threshold $\frac{1}{2} < r \leq 1$

such that any proposition A is believed if and only if the degree of belief in A is at least r. Formally: $Bel(A) \iff C(A) \geq r$. In a nutshell, Leitgeb shows that this can be done for the price of admitting that belief is *context-sensitive*, i.e., that the threshold value r, the credence function C, and the set of relevant propositions Π can be codependent. In his *stability theory of belief*, there then exists, in any context, a unique proposition B_W of stably high subjective probability ($C(B_W \mid A) > \frac{1}{2}$ for all compatible propositions A) such that $Bel(B) \iff B_W \subseteq B$.

For instance, one may care either about the event that participant 1 wins the lottery ($\Pi = \{\{w_1\}, \{w_2, \ldots, w_N\}\}$) or about the event that *someone* wins the lottery ($\Pi = \{\emptyset, \{w_1, \ldots, w_N\}\}$), but it is hard to imagine a realistic decision problem that would require a rational agent to represent both at once in the same logical algebra. The first partition Π is relevant to participant 1, who better act as if her ticket is not going to win (rather than buying a boat right away). The second is relevant to the lottery company, which better be prepared to pay out the jackpot.

The parallels between Cournot's principle and Leitgeb's *stability theory of belief* are fairly obvious, but the theory doesn't quite carry over one-to-one. It allows, in principle, for any value $r > \frac{1}{2}$, which is a bit too generous for appealing to CP. And in physics, we don't consider measures on any odd event space, but on the microscopic phase space of the theory. On the other hand, we are not interested in all implications (i.e., all directions of co-dependence) of Leitgeb's theorems, e.g., in first fixing what we believe and then looking for a consistent assignment of probabilities. Rational beliefs based on science should follow what the theory predicts, not the other way around. So while we cannot rely in the full authority of Leitgeb's proofs, his analysis carries over in the following sense:

In general, we do not care about *all* possible events $A \subset \Omega$ (at least not all at once). Especially in physics, when Ω is the microscopic phase space, most measurable subsets are just arbitrary collections of possible micro-configuration that do not correspond to any meaningful macro-event. Instead, a specific context of reasoning will be associated with a limited set of predicates ("macro-variables") Π_j partitioning Ω. And in each context, there can be another threshold value δ_j such that CP applies when

$$\mathbb{P}(A) > 1 - \delta_j, \; A \in \Pi_j. \tag{3.2}$$

In general, Π and δ can be "balanced" in such a way that this condition is closed under conjunction. And then, Cournot's principle applied to (3.2) will cohere with the Leitgeb's stablity theory of belief. In particular, there will be a strongest typicality fact entailing all others (see our Proposition 6.3.4).

Often, different contexts will be associated with different scales of time, area, and/or sample size. The probability of an earthquake with Richter magnitude ≥ 10 occurring next week in Sacramento is negligible (absent seismic indicators, cancelling a trip for fear of this event would be irrational); the probability of such an earth quake hitting California within the next 100 years or so is not. In general, different scales will be relevant for different epistemic agents, e.g., individual people as opposed to insurance companies or governmental regulators.

In a scientific context – when one could say that the relevant epistemic agent is a scientific community – we usually care about robust phenomena and statistical regularities. The class of phenomena that theories are tasked with explaining will, however, differ across the various disciplines, and the "threshold" for typicality seems to be different, in an interesting hierarchical way, for regularities falling under the purview of fundamental physics or the various special sciences. We will return to this point in Chapter 15.

Emile Borel (1939), discussing Cournot's principle (though not by this name), actually made a proposal for the orders of magnitude characterizing "negligible probabilities":

$p < 10^{-6}$ on the individual human scale

$p < 10^{-15}$ on the terrestrial scale

$p < 10^{-50}$ on the cosmological scale.

He argued: "In the ordinary conduct of his life, every man usually neglects probabilities whose order of magnitude is less than 10^{-6}, that is, one millionth, and we will even find that a man who would constantly take such unlikely possibilities into account would quickly become a maniac or even a madman." The spirit of Borel's reflections is more interesting here than the exact numbers.

3.2 The Rationality of Cournot's Principle

Does nature have to obey Cournot's principle?

We have introduced CP as a rationality principle rather than a factual claim, that is, as a statement about what we should expect rather than what will actually happen. Many would argue, however, that following CP can only be rational if it is successful; if, as a matter of fact, very improbable events occur at most as rare exceptions. But in what sense could the factual claim that a very unlikely event will not happen be implied by the normative claim that we should act as if it won't? In more metaphorical terms, the question is this: Who has to abide by Cournot's principle, rational agents or nature itself? There it is, the usual dialectic between the epistemic and aleatic nature of probability all over again.

Part of the synthesis lies in the process of *hypothesis testing*. We would not, and should not, accept a theory of nature if it makes the relevant phenomena unlikely. This is standard scientific practice: If the probability of an observation O under a hypothesis H is very low, we reject the hypothesis. Somewhat oversimplified, $\mathbb{P}(O \mid H)$ is the infamous p-value. The standard convention in special sciences sets the threshold for rejecting a "null-hypothesis" in a single study at $p = 0.05$ (and it is hotly debated at the moment, whether this value is too large); in particle physics, it is 5σ, or roughly $3{\cdot}10^{-7}$. But no matter how large the sample, or how often an experiment is reproduced, no p-value, however small, would make it *impossible* that a true hypothesis is falsely rejected (called a *type I error* in statistics).

The upshot is that rational scientists would never accept a theory that makes the phenomena atypical, even if that theory were actually true. But whether we can ultimately appeal to more than human rationality depends, to a large extent, on our metaphysical attitude towards laws of nature. The Humean best system account, which regards laws of nature as optimal summaries of contingent regularities in the world (see Chs. 5 and 13), can provide a very convenient answer: A true probabilistic law can get it wrong some of the time (i.e., assign a very low probability to an event that actually happens) but it cannot be wrong a lot of the times, or else it wouldn't be part of, or deducible from, the best systematization of the world. In my (anti-Humean) view, it is not a conceptual truth that the actual world does not correspond to an atypical instantiation of the laws, but a foundational belief of its scientific investigation. In this sense, I do believe that nature itself is bound by the rationality of Cournot's principle.

In the end, every good scientist accepts the rationality of the scientific method while also being aware of the possibility of error. For Humeans, this epistemic humility is only due to limited data (if she knew the entire mosaic, a sensible physicist couldn't be wrong about the laws). In my view, it is the right attitude towards the fundamental laws regardless of observational limitations. In practice, it won't make much of a difference, since we are, in fact, limited beings.

Moral certainty

Cournot's principle stands in the long philosophical tradition of *moral certainty*, which describes a degree of certainty that falls short of absolute metaphysical/logical/mathematical certainty but must nonetheless be considered sufficient for practical purposes or the purposes of a particular field of inquiry. This distinction goes back at least to Aristotle, who explains that it would be unreasonable to hold moral philosophy to the same standard of proof as mathematics (Nicomachean Ethics 1094b). Here, we can already see the double-meaning of moral certainty. In the more literal sense, it refers to the degree of certainty with which moral truths can be established. In the more relevant sense (for our purposes), it points to normative principles that compel us to accept certain inferences on pain of irrationality. This latter aspect comes out clearly with Leibniz, who writes:

> *Certainty* might be taken to be knowledge of a truth such that to doubt it in a practical way would be insane; and sometimes it is taken even more broadly, to cover cases where doubt would be very blameworthy (N.E. 445, quoted after Leibniz (1765/1982))

While Leibniz already invokes probabilistic notions, the connection to a mathematical theory of probability is first made explicit in Jakob Bernoulli's seminal *Ars Conjectandi* (1713) . Bernoulli defines something as "morally certain if its probability is so close to certainty that the shortfall is imperceptible" and "morally impossible if its probability is no more than the amount by which moral certainty falls short of complete certainty." He goes on to explain:

> Because it is only rarely possible to obtain full certainty, necessity and custom demand that what is merely morally certain be taken as certain. It would therefore be useful if fixed limits were set for moral certainty by the authority of the magistracy – if it were determined, that is to say, whether 99/100 certainty is sufficient or 999/1000 is required....

Both the pragmatic and the normative aspects of probability reasoning are discernible in this statement; so is the problem of vagueness, of fixing a threshold value for moral certainty/moral impossibility. Bernoulli wants to address this issue by appealing to "the authority of the magistracy" while we can only appeal to the authority of reason.

Black swans and Pascal's wager

> But that there is no science of the accidental is obvious; for all science is either of that which is always or of that which is for the most part.
>
> — Aristotle, Metaphysics (1027a)

While philosophers are prone to worry about the possible, no matter how implausible, overconfidence is the cardinal sin of the practitioner when it comes to probabilistic forecasts. Thus, it has been convincingly argued (e.g., in Taleb (2010); Silver (2012)) that many catastrophic prediction failures, in particular in economic, social, and environmental sciences, come from neglecting the possibility of low-probability but high-impact events (so-called "black swans"): market crashes, political revolutions, environmental disasters, etc. This important lesson may seem to go against the rationality of CP – the principle of the negligible event, as it has been called – as applied to singular events.

We have already made two salient remarks that can be repeated here. First, that pragmatic considerations may very well figure into choosing the threshold for "very low probability" and thus identifying the range of typical versus negligible outcomes. Second, that one has to identify the relevant events to begin with. Even the cunning investors who are successfully betting on rare events that the market tends to underestimate are following CP: they are betting on the typical *regularity* that unexpected market events happen occasionally.

I would insist on two further points. For one, that an individual acted irrationally by buying into the lottery even if he actually won the jackpot. This is to say that not all black swan events amount to a prediction failure challenging the rationality of Cournot's principle. Furthermore, the difference between a cautious forecaster and a paranoic is that the former will apply the principle of the negligible event *at some point*. If we could never neglect the possibility of extreme events on the basis of their minuscule probability alone, we would constantly have to buy into Pascalian wagers of sorts, since there is basically no upper bound on the magnitude of possible catastrophes.[1]

Finally, it may be interesting to note that exceptional events can seem much more significant in human affairs than in sciences tasked with investigating the regular course of nature. In particular, in the context of statistical mechanics, singular atypical events (one apple spontaneously jumping off the ground, let's say) would arguably appear like the kind of mid-size "miracle"[2] that would be dismissed as a false report or some sort of observation error but remain inconsequential for the discipline at large. In human affairs, on the other hand, it is the very unpredictability of "black swans" that tends to increase their impact.

[1] The ultimate "paranoid" scenario may be a decay of the (false) Higg's vacuum that would annihilate all the matter in the observable universe, see Mack (2015).

[2] Not literally a violation of the fundamental laws, though.

Chapter 4

A Typicality Theory of Probability

This section will finally introduce our proposed interpretation of deterministic probabilities as *typical relative frequencies*. In the language of typicality, this approach was laid out by Goldstein (2012) and further developed in the textbooks of Dürr et al. (2017) and Dürr and Lazarovici (2020, Ch. 3) that our discussion will partly follow. In addition, many landmark works in the history of probability could be claimed as precedents (see Ch. 3 on Cournot's principle) – one may even argue that Kolmogorov's axiomatization of modern probability theory was proposed with a similar view in mind – but it would be a matter of historical debate how far these claims can go.

Understandably, people who feel comfortable with the concept of probability are generally less sympathetic to our program, which seeks to ground it in – and in some instances replace it with – the new concept of typicality. That's fair. One just has to be careful not to confuse familiarity with comprehension, the successful application of the probability calculus with a clear understanding of what probabilities mean. The fact that the interpretation of probability has been controversially debated for centuries shows that there is at least no consensus, even when it comes to specific areas of application like statistical mechanics.

We consider the paradigmatic example of a "random experiment," which is the repeated tossing of a fair coin. A sequence of coin tossings, let's say of length $n = 1000$, can be viewed as a 0-1-sequence: 0 stands for *heads* and 1 for *tails* (let's say). Now consider the following mathematical facts:

1. The total number of possible 0-1-sequences of length 1000 is 2^{1000}.

2. The number of sequences of length $n = 1000$ with exactly k heads is $\binom{n}{k}$.

3. The values of this combinatorial factor for different k is shown in the following table 4.1. We care, in particular, about the relative number of sequences with $k \approx 500$ (equal distribution of 0 and 1) versus those with $k \ll 500$ (unequal distribution) .

k	**100**	**200**	**300**	**400**	**450**	**480**	**500**
$\binom{1000}{k} \approx$	10^{139}	10^{215}	10^{263}	10^{290}	10^{297}	10^{299}	10^{299}
$\binom{1000}{k}/2^{1000} \approx$	$\frac{1}{10^{161}}$	$\frac{1}{10^{85}}$	$\frac{1}{10^{37}}$	$\frac{1}{10^{11}}$	$\frac{1}{10^{4}}$	$\frac{1}{100}$	$\frac{1}{40}$

Table 4.1: Absolute and relative number of 0-1-sequences of length $n = 1000$ with k zeros. Note that $\binom{n}{k}$ is symmetric about $n/2$. The given values are approximate.

We note: $\binom{1000}{300}$ differs from $\binom{1000}{500}$ by a factor of 10^{36} (!). More generally, the number of sequences with a roughly equal distribution of 0's and 1's is *overwhelmingly greater* than the number of sequences with a distinctly uneven distribution. In fact, from the bottom line of the table we can readily estimate that sequences with $k \in [500 \pm 50]$ make up *nearly all possible ones*; sequences with fewer than 450 1's (or 0's) contribute almost nothing to the total number.

Thus, all sequences are different; some consist almost entirely of 0's, other contain twice as many 1's, etc. However, among the set of possibilities, we find a *typical regularity* that comes with large numbers (the "large number" here being $n = 1000$): in nearly all possible sequences, 0 and 1 (*heads* and *tails*) appear with roughly equal frequency. This equidistribution is typical.

Another typicality fact is that nearly all possible sequences are *irregular*. This is to say that they look nothing like 0101010101..., but that the occurrences of 0 and 1 seem unpredictable. There are different ways to make this precise (in terms of complexity, entropy, correlations, etc.) but a relatively simple argument is the following: Let's say we have roughly $2^6 = 64$ orthographical symbols at our disposal, including mathematical symbols. Then there are $\sum\limits_{k=0}^{m} (64)^k \approx 2^{6m}$ possible sentences of length m or less, compared to 2^n possible 0-1-sequences of length n, which is a far greater number if $n \gg m$. Therefore, the vast majority of sequences do not allow for any description that would be significantly shorter than the sequence itself.

In any case, some form of irregularity or unpredictability is characteristic of what we would call "random" behavior, and here we see that (apparently) random behavior is itself a typical phenomenon. Notably, this is true regardless of whether the fundamental process producing the sequence of events is deterministic or intrinsically random. A stochastic coin-toss law can produce very regular sequences like 0101010101... but only with very low probability (for large n). And a deterministic law can *typically* produce very irregular sequences that pass all statistical tests for randomness. (A concrete example will be discussed below.) This is also why the question of whether our world is, in fact, deterministic or indeterministic can never be settled on empirical grounds alone.

Irregular behavior in deterministic systems is closely related to the concept of *chaos* or *dynamical instability*. For actual coin toss experiments, we intuitively know what to do in order to produce "random" outcomes: the coin must be sent spinning and whirling, enough to make the result unpredictable because the smallest change in the initial (angular) momentum imparted to the coin by the throwing hand can lead to a different outcome, from heads to tails or vice versa. The motion of the coin becomes chaotic, in the sense that small causes can have large effects. This is very important for the appearance of random behavior, not just because of practical unpredictability but also because chaotic dynamics are usually linked with some notion of *statistical independence* that allows the law of large numbers to take effect.

4.1 Normal Numbers as a Model for Coin Tossing

Simple counting, as we just did to determine what is typical, won't do the job if we want to take the coin toss seriously as a physical process whose outcome is determined by dynamical laws and initial conditions. The initial conditions for the coin are positions and (angular) momenta, which are continuous variables, so we can no longer actually count the possibilities. But what can tell us now what is small and what is overwhelmingly large if there is an infinity of possibilities either way? The answer is a measure on the continuum, a *typicality measure*.

In a somewhat realistic physical analysis, this would be a measure on phase space – which is 12-dimensional if consider the coin as a Newtonian rigid body and roughly 10^{24}-dimensional if we consider it, from a microscopic point of view, as a collection of particles. Naturally, such an analysis is very difficult to do (virtually impossible in the microscopic case). Instead, we shall consider a mathematical model that is highly instructive and easy to analyze rigorously.

It is helpful for the discussion to avoid human involvement altogether and think of the coin being tossed, not by human hand, but by a coin-tossing machine, i.e., a mechanical device which takes a coin, tosses it, registers heads or tails, takes the coin again, tosses it again, and so on. Each time, the machine must flip the coin with slightly different angular momentum. But this in itself is no excuse to smuggle in probabilities since the machine itself is a physical system obeying deterministic laws. We shall imagine this system as isolated, that is, the machine is set up at some initial time t_0 with some initial condition x and from thereon everything runs like clockwork.

Let Ω be the physical state space of the machine together with the coin. The machine produces for each $\omega \in \Omega$ a sequence of coin-tossing outcomes, which is completely determined by ω. The results of each coin toss are given by *coarse-graining functions* on Ω, macro-variables mapping each $\omega \in \Omega$ to the value 0 or 1 (head or tails). To keep things mathematically tractable, we consider a simple model with $\Omega = [0, 1)$ and $\omega = x \in [0, 1)$. Hence, we look for functions that map the interval $[0, 1)$ to the value set $\{0, 1\}$ and capture the characteristic features of coin tossing, in particular the idea of *statistical independence.* The coarse-graining must produce this independence in interplay with a natural measure.

A big question when probability theory was being worked out as a mathematical discipline was: Are there any “natural” examples of such coarse-graining functions, or would one have to rely on contrived ad hoc constructions? The following realization, though somewhat forgotten in modern days, proved to be a huge stepping stone: Represent each $x \in [0,1)$ in binary form[1]:

$$x = 0.x_1x_2x_3\ldots, \quad x_k \in \{0,1\}, \;\; x = \sum_{k\geq 1} x_k 2^{-k},$$

so that $x_k \in \{0,1\}$ is the k-th digit in the binary expansion of the real number x. Now for $k \in \mathbb{N}$, consider the functions

$$r_k : [0,1] \to \{0,1\};\; x \mapsto x_k, \tag{4.1}$$

mapping the real number x to its k'th binary digit. These functions are called *Rademacher functions.* The first three are sketched in Fig. 4.1.

We can use the Rademacher functions to model coin tossing (cf. also (Kac, 1959, Ch. 2)). Think of x as representing the physical initial condition and of $r_k(x)$ as the outcome of the k'th coin toss given by the respective solution of the equations of motion plus coarse-graining (we care only about which side of the coin faces upwards). The role of the deterministic law is thus played by the binary expansion of the numbers.

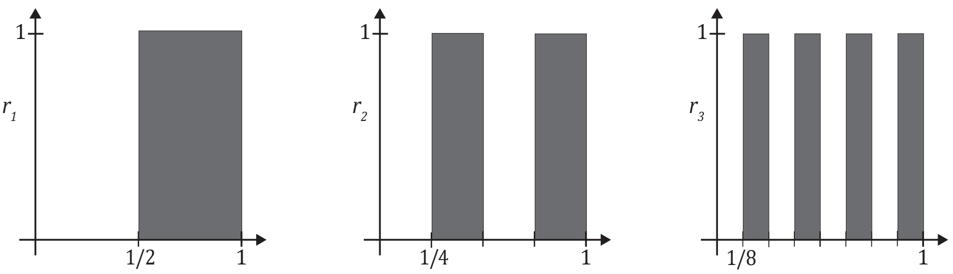

Figure 4.1: Area under the graphs of the Rademacher functions r_k for $k = 1, 2, 3$. The preimages are half-open intervals.

Remark (On the term “random variable”)**.** In the mathematical literature, such (measurable) coarse-graining functions are called “random variables.” This terminology is quite unfortunate (Mark Kac (1959) called it “horrible and misleading” (p. 22)), as it suggests something intrinsically chance-like about them. Here, it should be obvious that there is nothing chancy or indeterministic about the digits of a real number.

[1] With the convention that we write ...$1\overline{0}$ instead of ...$0\overline{1}$.

The Rademacher functions capture our intuitive understanding of independence. The values of $r_k(x)$ for $k \leq n$, i.e., the first n binary digits of x, tell us more about x (e.g., $r_1(x) = 1 \Rightarrow x \in [1/2, 1)$) but imply nothing about the $(n+1)$-th binary digit or any digit after that. The precise mathematical concept of statistical independence, however, is not a feature of the macro-variables or the dynamics alone but is defined in terms of a measure.

Coarse-graining functions, as the name suggests, are not one-to-one: Many different x are mapped by r_k to one and the same result. Which x values those are is given by the preimage $r_k^{-1}(\delta)$, $\delta \in \{0, 1\}$, for example

$$r_1^{-1}(0) := \{x \in [0,1) : r_1(x) = 0\} = [0, 1/2)\,.$$

Coarse-graining functions thus *partition* their domain into cells (the Boltzmannian “macro-regions” in statistical mechanics). In this particular case, the preimage sets of the values 0 and 1 have the same size or *content*. We have a pretty good intuition for what we mean by that: the content of an interval $[a, b)$ is its length $\lambda([a, b)) := b - a$, the content of a rectangle is its area, the content of a three-dimensional cuboid is its volume, etc. This leads to the construction of the *Lebesgue measure* λ on $\mathbb{R}^n$ and then to the abstract mathematical concept of *measures*, in general, but the prototype of all measures is the intuitive content.

In any case, with respect to this natural measure, it is easy to check that

$$\lambda\left(r_k^{-1}(\delta_k) \cap r_l^{-1}(\delta_l)\right) = \lambda\left(r_k^{-1}(\delta_k)\right) \cdot \lambda\left(r_l^{-1}(\delta_l)\right) = \frac{1}{4},\ \forall k \neq l,\ \delta_k, \delta_l \in \{0, 1\}. \tag{4.2}$$

This product structure (4.2) defines statistical independence. The Rademacher functions thus give us the necessary trust that statistical independence can indeed arise as a natural mathematical feature of coarse-graining functions. The independence of the Rademacher functions can almost literally be seen from Fig. 4.1. The coarse-graining yields a very distinct partition; the pre-images of r_k for different k mix or intertwine in an extremely orderly fashion as to realize (4.2). This is the ideal case, the paradigmatic example of statistical independence. In more realistic physical models, the “mixing” will be much messier and harder to picture, let alone prove.

Law of large numbers for Rademacher functions

With this groundwork, it is a standard mathematical exercise to prove a law of large numbers for the Rademacher functions. Let

$$m^n_{\text{emp}}(x) := \frac{1}{n}\sum_{k=1}^{n} r_k(x)$$

be the empirical mean[2], i.e., the relative frequency of 1's in the first n binary digits of $x \in [0,1)$. Exploiting statistical independence, one obtains:

$$\lambda\left(\left\{x \in [0,1) : \left|m^n_{\text{emp}}(x) - \frac{1}{2}\right| > \epsilon\right\}\right) \leq \frac{1}{4n\epsilon^2},\ \forall \epsilon > 0. \tag{4.3}$$

We can phrase this result in various ways:

- The set of $x \in [0,1)$ for which 1 and 0 does *not* appear roughly equally often in the binary expansion has *negligible content.*
- For the *overwhelming majority* of $x \in [0,1)$, the relative frequency of 1's and 0's is approximately 1/2.
- A relative frequency of $m^n_{\text{emp}} \approx \frac{1}{2}$ is *typical* in Ω.
- We can call this typical value $p = \frac{1}{2}$ of m^n_{emp} (its expectation value) the *probability* of "heads" (and, analogously, "tails"), which corresponds to the intuitive Laplace probabilities for coin tossing.

To clarify the transfer from this mathematical model to the relevant physical situation: Each $x \in \Omega$ corresponds to a possible initial condition or nomologically possible world instantiating a different outcome sequence of heads and tails which is completely determined by the dynamics. For some initial conditions, almost all the coins would land on *heads*; for others, *tails* comes out with much higher frequency. However, nearly all possible worlds instantiate the statistical regularity that the relative frequency of *heads* and *tails* in a long series of tosses is approximately 1/2. This (if it could be established) is an objective "non-random" fact about the possible worlds allowed by the deterministic laws, just as the typical distribution of binary digits is an objective fact about real numbers.

In standard textbook terminology, one would call

$$\mathbb{P}(k,\delta) := \lambda\left(r_k^{-1}(\delta)\right) \tag{4.4}$$

"the probability" of the outcome $\delta \in \{0,1\}$ in the k'th trial. But what would have been the point of using a word with so much philosophical baggage for something as pedestrian as the length of intervals? More generally, there are at least two reasons to avoid the identification of (4.4) with the physically relevant notion of probability.

1. While (4.4), technically an image measure" under the "random variable" r_k, is a useful and natural mathematical concept, it has no empirical content. What we observe and try to explain are statistical regularities, i.e., relative frequencies, not weights assigned to sets of possible initial conditions.

[2] A more accurate but somewhat unwieldy name would be "theoretical empirical mean" since m_{emp} yields the relative frequency as a function of the possible x rather than the one actually observed.

2. Many measures other than the uniform Lebesgue measure would make it true that $\mu\left(\left|m^n_{\mathrm{emp}}(x) - \frac{1}{2}\right| > \epsilon\right) \approx 0$ for large n, i.e., agree on the typical relative frequencies, while assigning different weights to the individual pre-images sets. In other words, probabilities understood as typical relative frequencies are very robust against variations of the typicality measure.

Symmetry and Laplace probability

We have, of course, a strong intuition about the tossing of a balanced coin. If the distribution of mass is symmetric between *heads* and *tails*, this should be manifested in an equal probability (corresponding to the *Laplace probabilities* of the "elementary events" *heads* and *tails*). Often, a principle of indifference is invoked to make this connection, as if the empirical regularity – the equidistribution of heads and tails in a long series of tosses – comes about because *we* have insufficient reason to prefer one side of the coin over the other. In fact, the connection between the symmetry of the coin, the symmetries of the physical laws, and the (roughly) equal frequencies of heads and tails is made by typicality. The physical symmetries are statistically manifested in typical "models" of the theory. In Ch. 5.5, we will see that a crucial argument for the "naturalness" of the uniform typicality measure on classical phase space is, in fact, not its uniformity per se but its invariance under the symmetries of Galilean spacetime, matching that of the dynamical laws.

Biased coins

Our first argument, based on a simple counting of possible outcome sequences, is vulnerable to an objection of circularity: haven't we essentially assumed an equal probability of *heads* and *tails* by counting all sequences with equal "weight"? For a biased coin, the set of possible outcomes would be the same, but the probabilities, i.e., the typical frequencies, should come out different from 1/2. Aside from the (crucial) fact that nothing about *probabilities* was *assumed* in our argument, this is a valid point.

One may now think about mounting a similar objection against the continuous model based on the Rademacher functions. Indeed, there are measures on the unit interval with respect to which "the great majority" of numbers would have a distinctly uneven distribution of digits. But those are measures that differ radically from the Lebesgue measure – measures that become *singular* in the limit $n \to \infty$ – and cannot be confused for sensible typicality measures. Mathematically, we could also put a delta-measure on $x = 0$ and say that almost all numbers are identically zero; but aside from technical jargon, this is simply an abuse of language and doesn't capture the correct notion of typicality.

So if probabilities other than 1/2 do not come from different typicality measures, where would they come from? Usually, from a difference in the *dynamics*, which partition the space of initial conditions (here $\Omega = [0, 1)$) into subsets leading to the outcomes

0 or 1, respectively. In our model, the dynamics were included in the Rademacher functions and the binary expansion of numbers which partition the unit-interval symmetrically. If we were to model an unbalanced coin, many initial conditions would be mapped to different outcomes.

Special "macroscopic" *boundary conditions* can also lead to different typical regularities. For instance, if we impose the boundary condition $x < 2^{-500}$ (so that $r_k(x) = 0, \forall k \leq 500$), the first $n = 1000$ digits of the binary expansion would typically contain about three times as many 0's as 1's. (Typically, that is, relative to the initial "macro-region" $[0, 2^{-500}) \subset \Omega$). This corresponds to what we call a *non-equilibrium* situation in statistical mechanics, and we emphasize that in the fundamentally deterministic setting, there are non-equilibrium macro-conditions rather than non-equilibrium typicality measures (here, the Lebesgue measure is always the appropriate typicality measure). In a very simplified sense, we can even see convergence to equilibrium in this model: the relative frequencies of 0 and 1 start out in non-equilibrium, with an over-population of 0's, and typically approach the equidistribution as n increases towards infinity.

4.2 Typical Frequencies

The theory that is beginning to emerge from our discussion relates probabilities to relative frequencies but is different from traditional frequentism or hypothetical frequentism. Slightly oversimplified, frequentists try to define the probability p of a (repeated) event as

$$p = \frac{1}{N}\sum_{i=1}^{N} \chi_i, \tag{4.5}$$

while hypothetical frequentists consider the limit of infinitely many (hypothetical) trials:

$$p = \lim_{N\to\infty} \frac{1}{N}\sum_{i=1}^{N} \chi_i. \tag{4.6}$$

The theory proposed here understands probability as *typical relative frequency*, that is

$$\mu\left(\left|\frac{1}{N}\sum_{i=1}^{N} \chi_i - p\right| > \epsilon\right) \approx 0, \tag{4.7}$$

where μ is a *typicality measure* and ϵ a small non-negative number.

Evidently, the conceptual distinction between typicality and probability measures (that we will further elaborate on) is essential here, otherwise an expression like (4.7) would be circular as a definition of probability. Probabilities, however, refer to statistical regularities, while the typicality measure is defined on sets of possible initial states and only used to identify which statistical regularities obtain for an *overwhelm-*

ing majority of initial conditions. Just as the Lebesgue measure on the set of reals, it has nothing to do with frequencies, or credences, or any kind of intrinsic randomness.

Indeed, among the many possible objections against finite and hypothetical frequentism – Hájek (1996, 2009) formulates 15 arguments against each – I want to highlight the following: Ultimately, any conclusive physical analysis has to speak about the universe and its initial conditions (cf. Ch. 8). And then, there simply are no meaningful frequencies, not even hypothetical ones.

Notably, (4.7) will, in general, not determine a unique number p but a small range $[p-\theta, p+\theta]$ of typical frequencies. It is only on the theoretical limit $n \to \infty$ that we can expect the typical value to become sharp. I consider this to be a feature, not a bug. The idea that probabilities correspond to exact real numbers is not born out by most interpretations and particularly naive under epistemic ones. At the very least, physical probabilities don't have to be sharp in order to guide rational credences, as credences are hardly determined up to infinitely many decimal places. They should, however, satisfy the axioms of probability in an appropriate sense.

Here, we find that typical relative frequencies are positive (since $\frac{1}{N}\sum_{i=1}^{N}\chi_i \geq 0$) and that the typical relative frequency of sure events is one (since $\frac{1}{N}\sum_{i=1}^{N} 1 = 1$). We must, however, impose the rationality principle that the reported typicality results should be as strong as possible, e.g., we should say that the probability of the sure event is one rather than "approximately 0.9999999" (which is technically true but silly).

Finally, typical relative frequencies for mutually exclusive events A and B are *additive* in the following sense[3]:

$$\begin{aligned} &\mu\left(\left|\frac{1}{N}\sum_{i=1}^{N}\chi_{A_i} - p\right| > \theta_A\right) = \delta_A \approx 0,\ \mu\left(\left|\frac{1}{N}\sum_{i=1}^{N}\chi_{B_i} - q\right| > \theta_B\right) = \delta_B \approx 0 \\ &\Rightarrow \left(\left|\frac{1}{N}\sum_{i=1}^{N}\chi_{A_i\vee B_i} - (p+q)\right| > \theta_A + \theta_B\right) \leq \delta_A + \delta_B \approx 0. \end{aligned} \tag{4.8}$$

If we called $(p-\theta)$ and $(p+\theta)$ the "lower," respectively "upper probability," this would be familiar from theories of *imprecise probabilities*. However, while most authors interpret imprecise probabilities subjectively, I submit that the objective physical probabilities that can be grounded in deterministic laws are (except in idealized limits) unsharp. Notably, the range of typical frequencies tends to get narrower with increasing sample size N (see our discussion of the LLN in Ch. 2). Physics usually deals with very robust phenomena (huge N) and thus very precise probabilities, while special sciences study regularities with much fewer instances and hence more imprecise probabilities. Notably, this is not just an epistemic or methodological claim but a physical one. The true probabilities (i.e., typical relative frequencies) that can be grounded in the fundamental laws of nature are much less sharp for macro-economic events than thermodynamic ones.

[3] This follows from the fact that $\chi_{A_i\vee B_i} = \chi_{A_i} + \chi_{B_i}$ for mutually exclusive events and thus $\left|\frac{1}{N}\sum_{i=1}^{N}\chi_{A_i\vee B_i} - (p+q)\right| \leq \left|\frac{1}{N}\sum_{i=1}^{N}\chi_{A_i} - p\right| + \left|\frac{1}{N}\sum_{i=1}^{N}\chi_{B_i} - q\right|$.

The significance of theoretical limits

In practice, the ensemble size N is not generally known or even fixed. What do we have in mind when we talk, for instance, about the propabilities, i.e., typical relative freuquencies, for coin tossing: $N = 1000$, $N = 1000000$, or maybe $N =$ #coin tosses actually occuring in our universe? If we are interested in precise estimates for the range of typical values, there is no way around specifying this (possibly alongside other relevant details of the coin-tossing process). For the colloquial use, however, it doesn't really matter. As long as N is reasonably large, the typical relative frequencies are approximately 1/2.

Could we also say that the typical relative frequencies for coin tossing are approximately $\frac{499}{1000}$ and $\frac{501}{1000}$? We could, but only when referring to coin-tossing sequences that are not too long, otherwise we might violate the principle that "the reported typicality results should be as strong as possible." The theoretical limit frequencies for $N \to \infty$, here 1/2, are thus distinguished by the fact that they are a good reference point for the range of typical values for arbitrarily large N. Following Goldstein (2012), we could call these limits *theoretical probabilities*, as opposed to physical probabilities, though only if this is not misunderstood as introducing two different philosophical concepts. The theoretical limits are just a convenient means to identify typical relative frequencies, both in the technical-mathematical sense – when it is useful to work with them – and in the pragmatic-linguistic sense – when we refer to the theoretical probability as representative for a small range of typical frequencies.

Speaking of these theoretical limits, we should not let equation (4.6) stand. From probability theory, a convergence of the empirical mean as in (4.6) can only be obtained as a typicality result. In particular, in the sense of the weak law of large numbers ("convergence in probability")

$$\forall \epsilon > 0 : \lim_{N\to\infty} \mu\left(\left|\frac{1}{N}\sum_{i=1}^{N}\chi_i - p\right| > \epsilon\right) = 0, \tag{4.9}$$

or in the sense of the strong law of large numbers ("almost sure convergence")

$$\mu\left(\lim_{N\to\infty}\frac{1}{N}\sum_{i=1}^{N}\chi_i = p\right) = 1. \tag{4.10}$$

But here, we are really in the regime of abstract mathematical models. Unless infinite ensembles are possible, there is no reference class of possible worlds to which such expressions successfully refer.

4.3 Probabilistic Predictions for Singular Events

To repeat: the role of the typicality measure in our account is not to assign an exact value to every possible event (or phase space region), but only to identify "very large" and "very small" sets of initial conditions. Probabilities, on the other hand, are taken to refer in the first place to statistical regularities rather than singular events. For this reason, I am not very concerned about the so-called *reference class problem* (see Hájek (2006) for a good discussion). One and the same event may be part of multiple statistical patterns, and which one we take to guide our actions or credences will usually depend on pragmatic considerations, not least on which typical regularities we are able to identify in the first place.

If I could see the initial orientation of a coin and be sensitive enough to narrow its initial angular momentum down to a small interval $[L_1, L_2]$, I might be able to place the respective coin flip in a statistical ensemble whose typical relative frequencies differ from $\frac{1}{2}$. Conceptually, this only poses a problem if one insists that there should be a physical probability associated with each *individual* coin flip event. There is, however, no contradiction between the statements "the typical relative frequency of *heads* in a long series of coin tosses is $\frac{1}{2}$" and "the typical relative frequency of *heads* in a long series of coin tosses with initial angular momentum $L \in [L_1, L_2]$ is $\frac{1}{3}$." Notably, no claim is made that using the additional available information leads to a more accurate prediction for that *particular* trial, but it will typically pay off in the long run.

Indeed, I find the interpretation of probabilities as typical relative frequencies compelling because single-case probabilities have no empirical content (a singular event either occurs or it doesn't). And I also see no reason why the fundamental laws of nature, if they are deterministic, should ground a probability for *any* odd macro-event, e.g., the outcome of the next presidential election.

One might worry, however, that typicality measures (as opposed to probability measures) are doing too little, by being, in effect, unopinionated about most possible events. The fundamental laws of physics may not predict probabilities (other than 0 or 1) for the next presidential election, but pollsters and political scientists certainly do. One could understand these probabilities epistemically. But if one holds a reductive view of special sciences (as I do, see Ch. 15), one must insist that if there's any objective sense in which such probabilistic predictions can be more justified or less, this should be at least partially grounded in physical facts. The following (idealized) example should be helpful to see how this could work.

Setup: We imagine an exit poll for a presidential election. Every single voter is assigned a number from 1 to N, and a pollster picks a sample of 1000 participants by a (classical) random number generator.

Result: 480 out of the 1000 participants respond that they have voted for the candidate of party A, while 520 respond that they have voted for the candidate of party B.

Fair sampling hypothesis: Each voter had an equal and independent chance of being polled. Hence, for each interview, the probability of picking a party A voter is equal to the actual share $p = \frac{k}{N}$ of votes that party A has received.

Mathematical fact: Under the above assumption, the sampling can be described as a Bernoulli process with unknown probability p. From the result of the poll, one can then compute a probability of approximately 0.8 that the actual percentage of party A voters lies below 50%.

Prediction: We should believe with 80% confidence that the candidate of party A has lost the popular vote.

Note that the probabilistic nature of this prognosis comes only from the "fair sampling hypothesis," not from an intrinsic probability of the predicted event. What the poll, together with the mathematical theory of probability, does, in effect, is to reduce rational belief about a complex event – the outcome of a presidential election – to rational belief about relatively simple events – the outputs of a random number generator. These beliefs about the sampling process can be physically justified by a typicality fact: it is typical that the number generator produces fair samples *in the long run*, i.e., all possible numbers with roughly equal frequency and in an unpredictable order. Yet again, there is nothing intrinsically random about this process (we are using a deterministic RNG), nor a physical probability associated with the election outcome per se.

This is all very basic statistics, and I am confident that someone more knowledgable in the field could provide a similar analysis for less simplistic scenarios. There is, of course, a bit of a "religious war" going on between Bayesians and Non-Bayesians, and I suppose the latter are the more natural allies for the typicality view. With respect to the former, the conjecture would be that all sensible priors would eventually convergence to the typical frequencies. To establish this rigorously, however, seems like a lot of work for another time.

Chapter 5

The Mentaculus

5.1 Typicality versus Humean Probability

Our discussion of the coin toss has already hinted at the general structure of deterministic physical theory, which is the following: We have a fundamental state space Γ whose points are the possible microscopic configurations (of the universe, in the last resort) and dynamical equations yielding a vector field on Γ which determines the possible time-evolutions of the states. The general solution of the equations of motion is described by a *flow* $\Phi : \mathbb{R} \times \mathbb{R} \times \Gamma \to \Gamma$ such that $X(t) = \Phi_{t,s}(x)$ is the unique solution with initial condition $X(s) = x$. Every $x \in \Gamma$ is thus a possible initial condition for the universe, corresponding to a possible micro-history.

In the context of classical mechanics, that we shall focus on for now, $\Gamma \cong \mathbb{R}^{3N} \times \mathbb{R}^{3N}$ would be the *phase space* comprising the positions and momenta of all the N particles, and the Newtonian or Hamiltonian equations of motion determine a Hamiltonian vector field. There is also a natural measure λ on Γ called the *Liouville measure*, which is just the Lebesgue measure, i.e., the intuitive phase space volume, in canonical coordinates. This measure has the special property of being *stationary* under (i.e., conserved by) the Hamiltonian dynamics, meaning $\lambda(\Phi_{t,s}A) = \lambda(A)$ for a (Borel) set $A \subseteq \Gamma$ and any $s, t \in \mathbb{R}$. In mathematical terms, the triple $(\Gamma, \Phi_{t,s}, \lambda)$ forms a *dynamical system.*

In anticipation of our detailed discussion of Boltzmann's statistical mechanics, we note that there is a reasonably widespread agreement that the following holds true as a mathematical fact (see, e.g., Bricmont (1995); Penrose (1989); Albert (2000); Carroll (2010); Goldstein (2012); Lazarovici and Reichert (2015)):

> There exists a small (low-entropy) region M_{PH} in the phase space Γ of the universe such that the uniform Liouville measure[1] λ on M_{PH} assigns high weight to initial conditions leading to micro-trajectories that instantiate the thermodynamic regularities – in particular, the *thermodynamic arrow of time* – and other salient patterns (about coin tosses, stone throws, etc.)

[1] If we can conditionalize on the constant total energy, the relevant measure is, more precisely, the induced *microcanonical* measure on the energy surface.

> that we observe in our universe.
>
> That is, if we denote this set of “good” initial conditions by $M^*_{PH} \subset M_{PH}$, it holds true that $\lambda(M^*_{PH})/\lambda(M_{PH}) \approx 1$.

Following recent lectures of David Albert, we shall call this the *fundamental theorem of statistical mechanics* (FTSM), although it is not literally a theorem in the sense of rigorous proof. “Statistical mechanics” here should be understood very broadly, as being tasked with explaining or predicting macroscopic regularities on the basis of the microscopic laws.

Some people will find it preposterous to refer to initial conditions of the universe in order to account for something like the motion of a rock or the cooling of a cup of coffee. Well, in practice, we don’t. In principle, however, even the best-isolated subsystem is part of a larger system with which it has interacted at some point. Hence, if we make postulates about initial conditions of various subsystems individually, we commit redundancy and risk inconsistency[2]. Any attempt at a conclusive and fundamental account must therefore talk about the universe as a whole.

An important question is, of course, why the mathematical statement seems so compelling, given that it is virtually impossible to prove for more than highly idealized models. This will be addressed in the next chapter, which discusses Boltzmann’s statistical mechanics in detail. Here, we shall focus instead on the physical and philosophical interpretation of the FTSM (assuming its truth), in particular the meaning and status of the measure figuring in it.

David Albert (2000, 2015) and Barry Loewer (2007a, 2012b) have developed a popular and well-worked out position in the context of the Humean best system account of laws (BSA), adapting David Lewis’ theory of objective chance (Lewis, 1980, 1994; Loewer, 2001, 2004). In a nutshell, the BSA regards laws of nature as the best systematization of contingent regularities; the “best” systematization being the one that strikes an optimal balance between simplicity and strength (informativeness) in describing the world (the so-called Humean mosaic). According to Albert and Loewer, the best system laws of our world consist in

1. The deterministic microscopic dynamics.
2. The Past Hypothesis postulating a low-entropy initial macrostate of the universe.
3. A probability measure $\mathbb{P} = \frac{\lambda}{\lambda(M_{PH})}$ on the Past Hypothesis macro-region M_{PH}.

This probability measure does not refer to any intrinsic probabilities or random events in the Humean mosaic. Its inclusion into the best systematization is justified by the fact that it comes at relatively little cost in simplicity but makes the system much more

[2] To adopt an expression from John Bell (2004, p. 166)

informative, precisely because it accounts – via the FTSM – for the thermodynamic regularities, the entropic arrow of time, and many other macroscopic phenomena.[3]

Loewer introduced the name “Mentaculus” for this best system candidate, a self-ironic reference to the movie *A Serious Man* in which a rather eccentric character tries to develop a “probability map of the entire universe.” As a philosophical proposal, though, the Mentaculus (eccentric or not) is certainly appealing, as it attempts to provide a precise account of objective probabilities in deterministic theories. Moreover, Albert and Loewer employ the Mentaculus in a sophisticated analysis of counterfactuals, records, and special science laws, the details of which are beyond the scope of this chapter.

The view defended in this thesis – and by other authors before (e.g., Goldstein (2012)) – is that the Liouville measure on the initial macro-region should be understood as a *typicality measure.* This is to say that the FTSM is interpreted not as a probabilistic statement but as the proposition that the macroscopic regularities in question obtain in *nearly all* possible worlds (consistent with the dynamical laws and the Past Hypothesis). In this sense, macroscopic regularities, such as the second law of thermodynamics, come out as *typical regularities* and the notion of “probability” is applied to describe typical statistical regularities but not to the fundamental measure on the phase space of the universe.

What makes the Mentaculus and the typicality account interesting subjects of a comparative analysis is that they agree on many basic points – like the objective nature of physical probabilities and the relevance of the FTSM – while disagreeing about three important issues:

i) What is the metaphysical status of (and justification for) the fundamental measure?

ii) What normative principle grounds its epistemic implications?

iii) And how wide is the scope of physical probabilities?

Therefore, it will be instructive to further develop the typicality view by contrasting it with the Mentaculus. A subsidiary goal of this section is to argue that it is preferable to adopt typicality even in the context of the best system account, while there are additional motivations if one holds an anti-Humean view about laws of nature.

5.2 Principal Principle and the Meaning of Humean Chances

We have already begun to explain that typicality facts are distinct from probability facts. In particular, that the role of a typicalility measure is only to determine “very

[3] It seems plausible that this basic structure of the best system does not hinge on particular micro-dynamics but can be maintained very generally; although when it comes to some of our current best theories, the details of the relevant probability postulate are far from clear.

large" and "very small" sets (of initial conditions, i.e., possible worlds) while no physical or epistemic meaning is attached to the exact number that it assigns to a particular set. The Humean probability measure, in contrast, is supposed to contain much more information. In fact, it will assign a probability (or conditional probability) to any physical proposition about the world: a probability that my dog gets sick if he eats a piece of chocolate, or that your favorite football team wins the next Super Bowl, or that the United States elect a female president in 2028.

When you ask a Humean to explain the regularity theory of chance in 5 minutes, you will hear something along the following lines: In our world, we find an irregular pattern of coin toss outcomes. Giving you a complete list of every single coin toss event would be very informative but not at all simple. Telling you that each coin lands either on heads or on tails would be very simple but not at all informative. Saying that the probability of *heads* and *tails* is 50% strikes the optimal balance between simplicity and strength. It summarizes the statistical pattern by saying that *heads* and *tails* come out in irregular order but with a relative frequency of 1/2 throughout the history of the world.

So far, so good, but in the Mentaculus theory Humean chances mean something different. First and foremost, the probability $\mathbb{P}(A)$ of an event A is the value that the fundamental probability measure $\mathbb{P}$ assigns to the set A of initial micro-conditions in the Past Hypothesis macro-region for which the respective event obtains (cf. Albert (2015, p. 8)). The epistemic and behavior-guiding function of these predictions is then supposed to be manifested in a normative principle, the *Principal Principle* (PP), which states that we should align our initial credences with the objective Humean chances. Formally:

$$C(A \mid \mathbb{P}(A) = x) = x, \tag{5.1}$$

or, for conditional probabilities,

$$C(A \mid B \wedge \mathbb{P}(A \mid B) = x) = x. \tag{5.2}$$

There are other variants of the PP proposed in the literature, and debates about what constitutes "admissible information" that one can conditionalize on (Hall (1994, 2004); Lewis (1994); Loewer (2004)), but these subtleties will not be relevant to our discussion.

In any case, stipulating the PP does not explain *why* it is rational to follow it, and what physical information a probabilistic prediction of the Humean best system contains. What exactly is the Mentaculus telling us by assigning, let's say, a (conditional) probability of 30% to the United States electing a female president in 2028? How, in other words, does the measure 0.3 assigned to the set of initial conditions evolving into a female president summarize what actually happens in our world? After all, the Lewis-Loewer theory agrees that there are no genuinely probabilistic facts in the mosaic. Every possible event either occurs or not, and whether it does is entailed by initial conditions and the deterministic dynamics. So what exactly are such single-case probabilities supposed to inform us about?

A standard Humean response is that, by definition, the probability measure figuring in the best system laws is the optimal measure for our world in terms of balancing simplicity and strength. Hence, while a single-case probability may not express anything about the individual event *per se*, there is something about the structure of the Humean mosaic as a whole that makes the particular value true or accurate (Lewis, 1980). Indeed, according to the BSA, $\mathbb{P}(A) = x$ is true in all and only those worlds whose best system implies $\mathbb{P}(A) = x$, so the proposition seems to be saying *something*.

I submit that we cannot get out more of the best system than we put in. The Humean probability law can only inform us about the features or regularities of the world that it is supposed to fit in the first place. If it is more accurate to assign a chance of 30% than of 60% (let's say), there must be concrete physical facts in the world that make it so; and these facts must be among those that go into evaluating the strength of the best system candidates. However, as I will argue in more detail below, many probability measures would assign a probability close to 1 to the thermodynamic regularities and other salient statistical patterns, yet a chance very different from 0.3 to the United States electing a female president in 2028. By some standard, these measures may not be as simple as the Liouville measure – which is why they are not part of the Humean best system – but unless they fare worse in terms of *fit*, there is nothing in the world that makes them less accurate when it comes to predicting presidential elections.

Some authors have read Lewis as suggesting that the Humean probability law is supposed to fit the macro-history of the world by assigning as high a probability as possible to any event that, in fact, occurs, and as low a probability as possible to any event that, in fact, does not occur (while being constrained by the requirement of simplicity). It cannot really work that way. In the competition for the best system, being a good predictor of presidential elections does not gain you as many points for "strength" as predicting the increase of the Boltzmann entropy in our universe. Also, assigning a probability of 1/2 to individual coin tosses – which may look like the laws are completely ignorant about the outcomes – is actually informative because it implies a very high probability for the event that the relative frequency of *heads* and *tails* in a long series of tosses is approximately 1/2. At the end of the day, the best system probability law will be one that informs us about robust regularities and global patterns in the world (by assigning to them a probability close to one), while the fit to singular events will count little to nothing in the trade-off with simplicity.

In particular: if, given the dynamical laws and the Past Hypothesis, two probability measures $\mathbb{P}$ and $\widetilde{\mathbb{P}}$ are equivalent in terms of strength – because they predict the same global patterns – while $\widetilde{\mathbb{P}}$ loses out in terms of simplicity, there is *no possible world* in which $\widetilde{\mathbb{P}}$ replaces $\mathbb{P}$ as part of the best systematization.[4] Therefore, a proposition like

[4] That is, unless the standard for simplicity is oddly contingent in a way that depends on microscopic

"according to the Mentaculus, the probability of event A is $\mathbb{P}(A)$ rather than $\widetilde{\mathbb{P}}(A)$" cannot restrict the set of possible worlds any further than to those instantiating the regularities on which $\mathbb{P}$ and $\tilde{\mathbb{P}}$ agree.

To sum it up in other words: According to the Humean view, there are certain "chancemaking patterns" Lewis (1994) on which a probability law supervenes, while "probabilities" for a great many other events – and, in fact, for all measurable subsets of phase space, most of which do not correspond to any meaningful macro-event – come out as a by-product. This by-product, to play on a metaphor by Albert (2015, p. 23), is not a gift from God ("I give you my most efficient summary of the regularities and you get rational credences for all conceivable events for free") but mostly mathematical surplus; the probabilities assigned to singular events could be very different, yet the physical content of the law the same. Sure, we can *postulate* that we should live our lives according to whatever numbers the Mentaculus spits out, but as so often with articles of faith, there is no rational prospect of reward in this world.

A true regularity theory of probability

In a nutshell, Humean probabilities are supposed to be efficient summaries of statistical regularities in the world. Then they turn out to refer, first and foremost, to a measure on sets of possible initial micro-conditions. What has one to do with the other? In most cases: nothing at all (is exactly my point). In some particularly nice cases, the connection between a statistical pattern instantiated by a series $S = (A_i)_{1\leq i \leq N}$ of similar events (e.g., a long series of coin tosses) and the probability $\mathbb{P}(A_i) = p$ of the individual events that make up the pattern (e.g., the i'th toss resulting in *heads*) is provided by a *law of large numbers* (LLN), that is, a result of the form

$$\mathbb{P}\left(x \in M_{PH} : \left| \frac{1}{N} \sum_{i=1}^{N} \chi_i(x) - p \right| > \epsilon \right) \propto \frac{1}{\epsilon^2 N} \approx 0. \tag{5.3}$$

Here, $\chi_i(x)$ is the indicator function mapping each possible initial micro-condition x to 1 if the event A_i occurs, and to 0 if the event A_i does not occur for the micro-trajectory with initial condition x (see Fig. 1).

We can read equation (5.3) as: "the measure of the set of initial conditions for which the relative frequency of occurring events deviates significantly from p is very close to 0." As a mathematical theorem, stating sufficient conditions for (5.3), the law of large numbers requires that the events are in some sense independent or uncorrelated, which is often intuitively compelling but nearly impossible to verify (and may fail much more often than we think). The standard proof of the LLN would furthermore make use of the fact that p comes out as the expectation value of the empirical distribution $\frac{1}{N} \sum\limits_{i=1}^{N} \chi_i(x)$. In the end, however, the role of the measure in (5.3) is merely to tell us that a particular set of initial conditions – the initial conditions that lead to

details or isolated macro-events.

significant deviations from the statistical pattern – has a measure close to zero. And at this point, it doesn't matter where the number p came from, whether it corresponds to $\mathbb{P}(A_i)$ or not, and whether we gave it any meaning as a probability in the first place. Its significance as a relative frequency describing a typical statistical pattern is established by, rather than assumed in, the law of large numbers result (5.3). Indeed, if (5.3) holds for the measure $\mathbb{P}$, an analogous statement will hold true for many other measures that agree on the smallness of that set (and, in general, of various other sets related to other statistical regularities), even if they disagree on the values assigned to events A_i individually. And it will also hold true in many cases in which the standard assumptions for the LLN are not satisfied. (For most physical applications, these are much too strong anyway, which is why statistical mechanics is so darn hard.)

Philosophically, it is thus unnecessary and misleading to think of (5.3) as a *consequence* of the single-case probabilities determined by $\mathbb{P}$. It is really the other way around: What a law of large numbers result does, in effect, is to reduce theoretical probabilities to *typical frequencies.* These typical frequencies, I claim, are all that the fundamental laws can or need to inform us about. In particular, if our best theory tells us that (with "near certainty") roughly 1/2 of the coin tosses result in *heads*, one can begin to justify the rationality of assigning credences about individual tosses accordingly; For instance by appealing to dutch-book arguments (if I accept bets of less than 2:1 on each one of these events, I can be almost certain to lose money on the long run) or maybe by invoking a principle of indifference with regard to the individual event in the pattern that we are about to observe (Schwarz, 2014).

Ultimately, this view is in no way tied to Humean metaphysics but when combined with a Humean regularity theory, it puts the latter back on its feet. As in the original 5-minutes sales pitch, probabilities are indeed referring to statistical patterns that the best system summarizes, rather than abstract weights it assigns to sets of possible initial conditions of the universe.

One may now wonder how this view of probabilities as typical relative frequencies makes sense of statements like: "The probability of event A: *My dog gets sick*, given B: *He eats the piece of chocolate I dropped on the floor* is p."

It seems natural to embed these events into a statistical ensemble: this and that fraction of dogs (of a certain size) get sick if they eat this and that amount of chocolate. If nature is kind to us, the best system will predict this statistical pattern as a typical frequency p – which may *or may not* correspond to the conditional probability $\mathbb{P}(A \mid B)$ that the Mentaculus assigns to the singular event that *my* dog gets sick after having swallowed the chocolate I dropped on the floor today. It also seems possible to decompose the events into a more fine-grained description which is part of a statistical pattern, e.g., the rate at which a dog's intestinal tract can metabolize theobromine or, finer still, interaction rates of certain molecules.

It seems to me that the intuition that singular macro-events could or should have an objective physical probability, in addition to a deterministic micro-description, comes from such (in principle) possibilities of embedding or decomposition – which are, notably, non-unique (recall the "reference class problem") and always require further context and analysis. It may be more difficult to explain in physical terms what one could mean by "the probability that the United States will elect a female president in 2028," but as a starting point, we should look at the statistical regularities, e.g., the sampling methods used to obtain polling data, that such predictions are actually based upon (see our example of the exit poll in the previous section). In any case, the idea that the Mentaculus provides a shortcut from the fundamental laws of physics to specific chance prescriptions for individual event tokens may be philosophically appealing but ultimately too simplistic to pan out.

In the end, whenever we succeed in grounding rational credences in the fundamental laws of physics, they will be grounded in *typicality facts*. A remarkable point worth emphasizing is that rational credences are thus grounded not in propositions about which our best theories are somehow noncommital or undecided, but in patterns and regularities that they predict beyond any reasonable doubt.

Principal Principle versus Cournot's Principle

What the previous discussion has, in fact, accomplished is to reduce the Principal Principle (PP) – at least those instances of the PP that could have a basis in physics – to a version of Cournot's principle (CP) that we discussed in 3. In contrast to the PP, we do not try to ground exact credences in any odd value that the fundamental measure assigns to any odd subset of phase space but care only about typical / atypical regularities. In the sense of Leitgeb's *stability theory*, we can also say that, in the context of statistical mechanics, the FTSM is the basic typicality fact which grounds or entails the typicality of various thermodynamic "laws" and statistical patterns. In any case, the measure of the "good" set of initial conditions is so close to 1 that its typicality is not in question.

While the Lewis-Loewer theory of objective chance is traditionally associated with the Principal Principle (Lewis even regarded PP as "non-negotiable"), it is very much, if not more, compatible with Cournot's principle: If the Humean probability of an event is very close to one, we can be almost certain that this event actually occurs. Why? Because this is what the best system is trying to tell us; because the way in which the Mentaculus summarizes relevant regularities in the mosaic is by assigning to them a probability very close to one. Ironically, a version of what Lewis (1994) considered to be the "big bad bug" of his theory of objective chance can serve to vindicate even the strongest form of CP in some cases. The Mentaculus will assign a very small though positive probability to the universe evolving on an entropy-decreasing trajectory. However, if the universe actually did evolve on such a trajectory, this Mentaculus would not be the best systematization of our world (given that so many salient features depend on its entropic history). Hence, the fact that the Mentaculus assigns a near-zero probability to anti-entropic trajectories, together with the fact that the

Mentaculus is the best systematization of our world, implies that an anti-entropic evolution of the universe is impossible. On the other hand, if we are talking about an event that the best system could, in principle, fail to fit, it is rather immaterial if it assigns a probability of 10^{-50} or 10^{-100}. Our residual uncertainty about whether that event obtains after all, does not come from anything the best system tells us about the world, but from the possibility that it just had to get this one wrong in the trade-off with simplicity.

In any case, our main argument so far why even Humeans should favor CP over PP is that the concrete physical information which the Mentaculus provides is to be found, first and foremost, in statements of probability close to 1 and 0, while the rationality of aligning credences with any odd value of the Humean probability is spurious, at best. Note that also methodologically, the only way to *test* probabilistic laws is by applying Cournot's principle, i.e., by rejecting the law-hypothesis if we observe phenomena to which it assigns a negligibly low chance (cf. Shafer and Vovk (2006)). Since I am not a verificationist, I do not claim that single-case probabilities are meaningless just because they cannot be empirically tested. I have, however, argued that the Humean regularity account fails to give them meaning as deterministic chances – except to the extent that they can be reduced to typical frequencies. Humeans often claim that their probability measure is "empirical," yet provides information far beyond what is empirically testable. I don't think they can have it both ways.

5.3 Probability versus Typicality Measures

The next step from probability to typicality comes by emphasizing the following insight: If we agree that all we need on the fundamental level are "probabilities" close to 1 and 0, then a whole lot of different measures could do the job.[5] If we don't like the Lebesgue measure, how about putting a (truncated) Gaussian measure on M_{PH}? In fact, we can tweak the measure in almost any way we like. Any measure that doesn't differ radically from λ will make a statement analogous to the FTSM true, and thus imply the same thermodynamic laws and statistical regularities. (In Chapter 8, we will discuss in more mathematical detail how the difference between two measures can be quantified.) We cannot be too extreme, of course. A delta-measure concentrated on an anti-entropic microstate will, evidently, lead to very different predictions. However, as Maudlin (2007b, p. 286) concludes: against this backdrop, "our concerns about how to pick the 'right' probability measure to represent the possible initial states ... or even what the 'right' measure means, very nearly evaporate."

An important observation is that probabilities (or weights, to use a more neutral term) close to 1 or 0 are very robust against variations of the underlying measure. Think for instance of the normalized Liouville measure as a uniform density over the

[5] The insight that a great many probability distributions over initial conditions lead to the same statistical predictions is also discussed (though from a somewhat different perspective) in Myrvold (2016) and, with more historical background, in von Plato (1994).

macro-region M_{PH}. If $\lambda(A) \approx 0$ but $\mu(A) \gg 0$, then the measure μ must radically differ from λ on a small set A. (The same holds by contraposition for probabilities close to 1.) In contrast, for $\lambda(B) = 0.3$ while $\mu(B) = 0.4$ (let's say), μ needs to deviate only mildly from the uniform density on the larger set B (see Fig. 5.1). "Large" and "small" is here understood with respect to the Liouville measure, but this doesn't make the argument circular. The point is that radical deviations from the Liouville measure would be necessary to come to different conclusions about typical/atypical regularities, while relatively small variations can lead to significantly different probability assignments for other events.

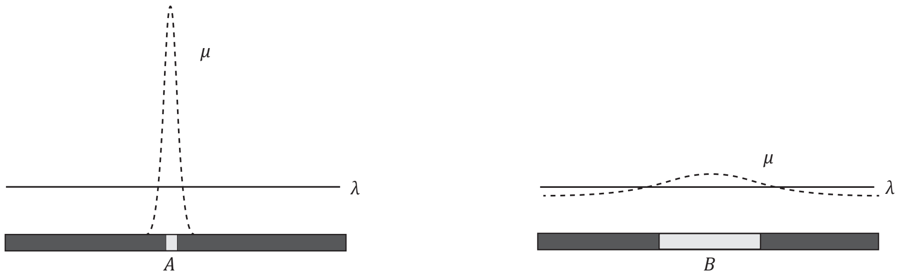

Figure 5.1: Schematically: λ and μ disagreeing significantly on the weight assigned to a set of small (left), respectively medium (right) λ-measure.

At least for the sake of argument, Albert and Loewer are willing to concede that we could consider best system candidates that involve an entire set or equivalence class of probability measures, with the understanding that the theory endorses all and only those probability statements on which these measures (more or less) agree (see, e.g., Albert (2015, footnote 2)). However, there is no good motivation to do so from a Humean perspective since a set or equivalence class of measures is neither simpler nor more informative than the Liouville measure (I have only argued that it is equally informative). The situation is rather the following: we should use the Liouville measure because it is simple and natural, but with the understanding that it is not a bona fide probability measure but a *typicality measure*. Its role and purpose is to designate events as "typical" (measure ≈ 1) or "atypical" (measure ≈ 0) or neither, while the precise numerical assignments have no physical meaning.

This move, to relate CP to a concept of typicality rather than probability, is a more recent development, though there is precedent for it in the physical literature (see, e.g., Everett (1973) and the discussion in Barrett (2016); also Bell (2004, Ch. 15), originally published in 1981, Dürr et al. (1992), and Goldstein (2001) on Boltzmann). There are additional motivations for this step, some of which we have mentioned before:

1. As argued above, Cournot's principle suggests an understanding of probabilities as typical frequencies. Of course, such an account would be circular if "typical" were itself explicated in probabilistic terms.

Moreover, probabilities can be understood as typical frequencies when applied to events *in* the universe, but we better not refer to frequencies (not even hypothetical ones) when we speak about the universe itself. For this reason alone, it makes sense to distinguish two different concepts.

2. In statistical mechanics, it is common and convenient to formalize typicality with the mathematical tools of measure theory, hence the deceptive kinship to probability. There are, however, other ways to define "typical," e.g., in terms of cardinalities of sets or dimensions of subspaces, which can be relevant in other contexts and figure in the same way of reasoning. (We will discuss this in more detail in Ch. 6; for more subtle technical differences between typicality and probability, see Wilhelm (2019).)

3. Some authors have suggested that "typical" is a more intuitive and unambiguous notion than "probable" (see, e.g., Dürr et al. (2017)). One way to spell this out is to say that the intuition associated with a typicality measure is one of "large" versus "small" rather than "probable" versus "improbable" sets and that we have a better intuitive grasp of the former than the latter.

   Another way is to compare the following two formulations of Cournot's principle:

   i) Expect to find the regularities that obtain in nearly all nomologically possible worlds.

   ii) Expect to find the regularities that obtain with probability close to 1.

   I would argue that the first rationality principle has an immediate intuitive appeal while the second version seems more stipulative, or at least neutral with regard to the interpretative question, what fact the probability statement actually expresses. Moreover, ii) can be applied to any probability measure – and has meaningful content only in conjunction with a particular measure – while i) has intuitive appeal only to the extent that the measure on possible worlds captures the intuitive meaning of "nearly all."

5.4 The Epistemic and Metaphysical Status of Typicality

The last points are, admittedly, controversial. Even some proponents of typicality are uncomfortable with the idea that there exists an a priori notion of "typical." And advocates of the Mentaculus emphatically deny that there are typicality facts that come more or less for free once the dynamical laws are fixed. The deeper question here concerns the metaphysical status of the measure. According to the Mentaculus account, the fundamental probability measure has the same status as any other Humean law: it supervenes on the contingent regularities in the world as part of the best systematization. This option is, in principle, also available for the typicality measure; that is, one could consider a typicality rather than a probability measure as part of the

Humean best system (Callender, 2007). In my view, however, the Humean account – regardless of its flaws or merits as a metaphysics of laws, in general – fails to capture the more subtle aspects of typicality and its use in physics. These come across if one considers the typicality measure[6] not as another theoretical postulate on par with the dynamics but as a *way of reasoning about* the dynamical laws. In other words: being intimately tied to Cournot's principle – which is normative rather than descriptive – the typicality measure falls itself, at least in part, into the normative domain.

Let me start to explore this by mapping out some key metaphysical differences between a typicality measure and a Humean probability measure:

1. According to the Mentaculus account, the Humean mosaic is the truthmaker of the probability measure as part of the best system.

   According to the typicality account (at least the version I am defending), a choice of typicality measure can be *reasonable* or *justified*, but there are no concrete physical facts that make it, strictly speaking, true. (I am sympathetic to the view that there are objective *normative* facts that make it true but that's beyond the scope of this discussion.)

2. According to the Mentaculus account, the probability measure, together with the other best system laws, cannot be entirely and radically wrong about the world, or else they would not form the best system.

   According to the typicality account, it is logically and metaphysically *possible* for a world to be – in any and all relevant regards – atypical with respect to the reference class of nomic possibilities.

3. Humean probabilities are supposed to summarize regularities in the actual world.

   Typicality statements summarize the modal structure of the laws. They do not refer directly to the actual world but to the fact that a certain feature or property is typical among all nomologically possible ones.

All these points show that the typicality measure cannot be just another bookkeeping device for the mosaic in the sense in which Humeans want to conceive of laws of nature in general. It can and should be tied to the dynamical laws (which, in turn, may be reducible or irreducible) but not in a way that depends substantially on contingent features of the world.

More precisely, I do not claim that the correct typicality measure follows deductively from the micro-dynamics but argue that there is no possible world in which the dynamical laws are the same as in ours, while the notion of typicality is significantly different. Imagine, for instance, a world that is consistent with Newtonian dynamics but anti-entropic micro-conditions, a world, that is, in which apples sometimes jump spontaneously up in the air, in which gases tend to clump, and gravitating systems to

[6] I am using the singular, though we should keep in mind that many different measures, qua mathematical objects, are equivalent as typicality measures.

be blown apart. I do not mean a world that is an exact time-reversal of ours (which would really be the same world if there is no primitive direction of time) but one in which violations of our second law of thermodynamics occur on a regular basis. It seems evident to me that the best system of this world – if one exists – would not involve Newtonian gravity with a strange typicality measure but very different micro-dynamics in the first place.

Or, if I can stop pretending to be Humean for a second: if the laws instantiated in that world were, in fact, the Newtonian ones, rational physicists would justifiably come to wrong conclusions about them. They would not discover the true dynamical laws but have a radically different understanding of "typical" than we do. Typicality, at least when combined with an anti-Humean conception of laws, thus allows for the metaphysical possibility of "unlucky suckers," epistemic agents in the absurd situation that rational inferences lead to radically wrong conclusions about their world. Conversely, to the extent that the laws of nature are epistemically accessible to us, our world must correspond – in the relevant respects – to a typical model of the true theory. This is nothing we could ever know with metaphysical certainty; we can only trust that "God is subtle but not malicious", as Einstein put it.

As usually, though, the epistemic situation for the Humean and anti-Humean is the same in practice. While I believe that it is always the theoretical system as a whole that is challenged by empirical evidence, I cannot conceive of a situation in which it would seem rational to revise our notion of typicality instead of adjusting the dynamical postulates (while the converse is common and unproblematic). For instance, there are almost certainly initial conditions for a Newtonian universe which are such that particles create an interference pattern whenever they are shot through a double-slit and recorded on a screen. This and other quantum phenomena are not made *impossible* by classical mechanics, they just come out as *atypical.* However, changing the typicality measure in order to save Newtonian theory from falsification by quantum phenomena is not a serious scientific option that anyone has ever, or should ever, entertain.

Using Quine's picture of a "web of belief" (Quine, 1951), I suggest that the dynamical laws are closer to the edges of the web than our notion of typicality. The typicality measure is somewhere in between the dynamical laws and the logical inference rules. This squares the circle between its being *necessary* but *not a priori.* While it is in principle possible to adjust it to new empirical evidence, the typicality measure is never the first knob to turn before making adjustments in other parts of a theoretical system.

One reason is that, because the notion of typicality is so robust against variations of the measure, any revision of it would be radical, i.e., would have to correspond to extreme changes in the measure on the state space of the theory. While it can be maintained that the change of the dynamics from Newtonian mechanics to quantum mechanics (or general relativity, etc.) was radical, as well, the new laws do at least recover the old ones in relevant limiting cases. It is hard to see how a similar continuity could hold between different notions of typicality. Instead, we would have to accept that we have been radically wrong about the meaning of "nearly all."

There is a more important reason why the typicality measure *should* be less empirical or epistemically more robust than the dynamics. As mentioned in the introduction, due to the huge number of microscopic degrees of freedom, the dynamical laws put barely any constraints on what is physically possible on macroscopic scales. Given any macroscopic phenomenon, there are almost certainly microscopic initial conditions for which the laws of motion would entail it. By the same token, given almost any micro-dynamics and any phenomenon in the world, there will be some measure that makes the phenomenon "typical" or sufficiently likely.

Therefore, treating the typicality measure on the same footing as the dynamical laws would give us too many moving parts that can be adjusted to fit the data. For the Humean, this is bad because it increases the risk of a tie for the best system, at least in hypothetical situations when simplicity of the dynamical and probability postulates pull in opposite directions. For the anti-Humean, it is even worse since the more freedom we have to adjust what counts as "typical" and "atypical," the less can empirical evidence inform us about the true laws of our world. To put it differently: the more we regard the typicality measure as physically contingent, an independent empirical hypothesis, the less explanatory work is done by the dynamical laws.

Example. Consider the following model for the coin toss, analogous to our discussion in Ch. 4. Let $\Gamma = (-1,1) \subset \mathbb{R}$ be the microscopic state space, respectively the relevant initial macro-region. Let $\Gamma \ni x = \pm \sum\limits_{k=1}^{\infty} r_k(x)\, 2^{-k}$, i.e., $r_k(x) \in \{0,1\}$ is the k'th digit in the binary expansion of x. We interpret the Rademacher function r_k as a macro-variable describing the outcome of the k'th coin toss. As discussed in 4, it is straightforward to show:

$$\lambda\left(\left\{x \in \Gamma : \frac{1}{N}\Big|\sum_{i=k}^{N} r_k(x) - \frac{1}{2}\Big| > \epsilon\right\}\right) \leq \frac{1}{4N\epsilon^2}. \tag{5.4}$$

Read: for large N, typical initial conditions w.r.t. the Lebesgue measure λ are such that the relative frequency of heads and tails is approximately $\frac{1}{2}$.

However, now consider the following family of truncated Gaussians as probability measures: For $n \in \mathbb{N}$, let $\mathcal{N}(0, \sigma^2(n))$ be the normal distribution with mean 0 and standard deviation $\sigma(n) := \frac{1}{10}\frac{1}{2^n}$ and

$$\mu_n := \frac{\mathbb{1}_{(-1,1)}\,\mathcal{N}(0,\sigma^2(n))}{\|\mathbb{1}_{(-1,1)}\mathcal{N}(0,\sigma^2(n))\|_1}, \tag{5.5}$$

where $\|\mathbb{1}_{(-1,1)}\mathcal{N}(0,\sigma^2(n))\|_1$ is the normalization. Then, almost the entire weight of μ_n (more than 10σ) is concentrated on the interval $I(n) := \left(-(\frac{1}{2})^n, (\frac{1}{2})^n\right)$, on which

$r_k(x) = 0, \forall\, k \leq n$. Hence, μ_n makes it overwhelmingly likely that the first n coin tosses result in *tails.* We could thus fit the statistical regularity without revising the dynamics – even without making the probability measure less simple – no matter how dominant the occurrence of *tails.*

In scientific practice, typicality judgments *do*, in fact, have a privileged status that the Humean regularity theory fails to account for.[7] In particular, atypicality is precisely the standard by which dynamical theories are reasonably rejected as empirically inadequate (think again of the double-slit experiment as a falsification of classical mechanics or the 5σ-standard commonly used in particle physics). Interestingly, this applies in pretty much the same way to deterministic laws as to intrinsically stochastic ones. The difference is that, in the latter case, falsifying a dynamical and a probabilistic hypothesis is one and the same, while for deterministic theories, it is primarily the dynamical postulates that stand trial. This is only possible because the typicality measure is epistemically more robust or, in some sense, entailed by the dynamical laws. In other words: in the scientific enterprise, some concept of typicality and atypicality is part of the backdrop against which law-hypotheses are evaluated rather than another law-hypothesis in its own right.

5.5 Justification of Typicality Measures

For these reasons, most advocates of typicality do not consider the typicality measure as an independent postulate of the physical theory, although it might be from a strictly logical perspective. But what then determines the right measure and accounts for its epistemic rigidity?

One answer we have already alluded to is that the role of the measure is not so much to *define* "typical" but to formalize an intuitive and largely pre-theoretic notion. In other words, there aren't competing versions of typicality corresponding to different choices of typicality measures, but one unified concept that a measure can either capture or fail to capture. If the set of possible worlds were finite, we wouldn't feel the need for an additional postulate to express what we mean by "nearly all possible worlds." In the continuum case, there is more ambiguity about how to "count," but this is arguably a technical rather than a conceptual issue.

The concept of typicality has, in any case, a certain vagueness. Just as it is impossible, in general, to specify a precise threshold ratio above which we should say that a subset contains "nearly all" elements – i.e., a precise threshold measure close to 1 above which something counts as "typical" – it seems impossible to specify a fixed set of criteria that make a measure convincing as part of the mathematical formalization

[7] Marc Lange (2009) makes a similar point when he argues for "degrees of necessity" in laws, though typicality is not quite a law, and nomic necessity is not quite the right concept here.

of typicality. In the context of classical Hamiltonian mechanics, the Liouville measure is clearly a reasonable choice while a delta-measure is clearly not, but a certain grey area in between seems unavoidable. Consider, for instance, a family of Gaussian measures with standard deviation $\sigma \to 0$ so that the distributions become more and more peaked (see the example above). Again, it would be misguided to ask for a sharp threshold value of σ below which the Gaussians cease to be suitable typicality measures. Still, this doesn't mean that the concept itself is ill-conceived or that the vagueness is problematic in practice. The bottom line is that a typicality statement has normative implications if and only if it is made with respect to a reasonable notion of "large" versus "small" sets. And while it is hard to state mathematically what makes a measure reasonable or unreasonable, we can generally tell them apart when we see them.

Some authors put less emphasis on the intuitive content of typicality and more on the condition that the measure must be *stationary* under the dynamics. This is to say that the measure of a set of microstates at one time must correspond to the measure of the time-evolved set at any other time. In this way, the dynamical laws themselves would constrain the choice of typicality measure. I will provide a more precise definition of stationarity and further justification for this condition below.

Fortunately or unfortunately, as long as we are dealing with classical mechanics, both approaches lead to the same conclusion since the simplest and most intuitive measure – the uniform measure on phase space – is also stationary under the Hamiltonian dynamics (though not uniquely so). It could even be justified by a principle of indifference (Bricmont, 2001), although I believe that the epistemic connotations are doing no good.

There is, however, a notable example where stationarity and intuitiveness seem to go apart (and the principle of indifference to fail altogether). In Bohmian quantum mechanics, the natural typicality measure grounding Born's rule and thus quantum statistics for subsystems is given by the $|\Psi|^2$-density on configuration space, induced by the universal wave function Ψ (Dürr et al. (1992), reprinted as Ch. 2 in Dürr et al. (2013b)). This measure is stationary (more precisely: equivariant) under the Bohmian particle dynamics and even uniquely determined by this condition (Goldstein and Struyve, 2007). However, since we do not know what the universal wave function and hence the induced typicality measure looks like, it appears that its justification can hardly lie in pre-theoretic intuitions. If the $|\Psi|^2$-density turned out to be sharply peaked, it would look very different from a uniform measure and thus seem to constitute a radical departure from the notion of typicality employed in classical mechanics. In the next subsection, I will discuss how this conclusion can be avoided and the two justifications of the typicality measure reconciled, after all.

Stationarity, uniformity, symmetry

The following discussion is quite technical, though the basic point is rather simple: So far, we have mostly talked about measures on the initial macro-region of the universe. In fact, the relevant reference class for typicality statements – what we actually want to quantify – are not initial micro-conditions but (nomologically) possible worlds. Initial conditions are just a means to parametrize the respective solution trajectories. Stationarity, simply put, guarantees that a large set of solution trajectories is deemed large by the same measure if we look at the trajectories, i.e., the corresponding set of microstates, at any other time.

One subtlety that often leads to misunderstandings is that proponents of the Mentaculus tend to think of the fundamental probability measure on phase space Γ as the one which is uniform over the Past Hypothesis macro-region $M_{PH} \subset \Gamma$ and zero outside. This is not a stationary measure on Γ since weight will "flow out" of M_{PH} and disperse all over phase space. Other authors tend to think of the stationary Liouville measure on all of phase space as the natural typicality measure which is then *conditionalized* on the initial macrostate M_{PH}. Notably, in our further considerations which focus on the solution space rather than phase space, this difference will become largely immaterial.

Let $\mathcal{S}$ be the set of solution trajectories for the microscopic dynamics (consistent with the Past Hypothesis) in the state space $\Gamma \cong \mathbb{R}^n$. For any $t \in \mathbb{R}$, let $\varepsilon_t : \mathcal{S} \to \Gamma, X \mapsto X(t)$ be the map evaluating the trajectory X at time t. These maps can be read as charts, turning the solution set $\mathcal{S}$ into an n-dimensional differentiable manifold.[8] The transition maps between different charts are then $\varepsilon_t \circ \varepsilon_s^{-1} = \Phi_{t,s}$, where $\Phi_{t,s}$ is the flow arising as the general solution of the laws of motion (Fig. 2).

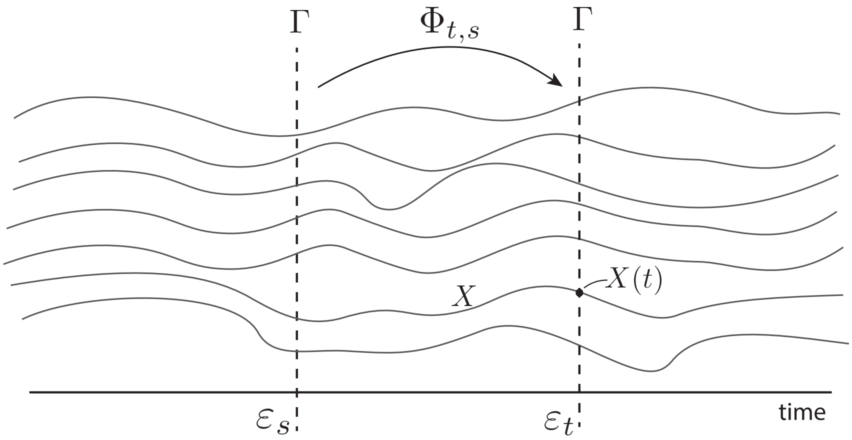

Figure 5.2: Sketch of the solution space and its parameterization by time slices.

[8] Strictly speaking, some solutions may exist only on a finite time-interval so that the charts are only locally defined. Here, we assume global existence of solutions for simplicity.

Now, the easiest way to define a measure μ on S is in one of these charts, let's say ε_0. Indeed, a possible point of view is that there exists a distinguished "initial" time so that it makes sense to parametrize solutions by initial data at $t = 0$. The other point of view, emphasized by our geometric notation, is that the choice of the "time slice" is arbitrary, essentially amounting to a particular coordinatization of the solution space. Under a transition map (coordinate transformation), the measure transforms by a pullback, $\mu_t = \Phi_{t,0}\#\mu_0$, i.e., $\mu_t(A) = \mu_0(\Phi_{t,0}^{-1}A)$ for any measurable $A \subset \Gamma$, where μ_t is the measure represented in the chart ε_t.

A measure is *stationary* if and only if it has the same form in every time-chart, i.e.,

$$\mu_t = \Phi_{t,0}\#\mu_0 = \mu_0, \; \forall t \in \mathbb{R}, \tag{5.6}$$

Equivariance is the next best thing if the dynamics are themselves time-dependent. Concretely, in the case of Bohmian mechanics, the particle dynamics are determined by the universal wave function Ψ_t which itself evolves in time according to a linear Schrödinger equation. Nonetheless, we have

$$|\Psi_t|^2 \mathrm{d}^n x = \Phi_{t,0}\# \left(|\Psi_0|^2 \mathrm{d}^n x \right), \; \forall t \in \mathbb{R} \tag{5.7}$$

so that the measure has the same functional form in terms of Ψ_t for any time t.

In conclusion, a stationary or equivariant measure on the state space Γ induces a canonical measure on the solution space $\mathcal{S}$: a measure that can be defined without distinguishing a set of coordinates, i.e., a particular moment in time.

Uniformity of a measure, on the other hand, is a *metric* notion. It requires that

$$\mu(B(x,r)) = \mu(B(y,r)), \; \forall x, y \in \Gamma, r > 0, \tag{5.8}$$

where $B(x,r)$ is the ball of radius r around x. However, even if the state space Γ comes equipped with a metric, it does not, in general, induce a canonical metric on the solution space $\mathcal{S}$. It is thus not clear what the geometric distance between two possible worlds should be, or whether it makes sense to regard $\mathcal{S}$ as a metric space (Riemannian manifold) at all. However, without a metric on the solution manifold, it is meaningless to ask whether a measure on it is uniform or not. From this point of view, it is indeed misleading to regard uniformity as a criterion for the typicality measure. Even the Liouville measure in classical mechanics is uniform on the "wrong space," namely on phase space rather than the space of possible worlds.

The uniformity of the Liouville measure on phase space does nonetheless capture a meaningful and important feature of the typicality measure, namely its *invariance under Galilean symmetries*. A symmetry is an isomorphism $T : \Gamma \to \Gamma$ that commutes with the flow, i.e., $\Phi_{t,s}(Tx) = T\Phi_{t,s}(x)$. This then induces a transformation $T^* : \mathcal{S} \to \mathcal{S}$ on the solution space by $T^* = \varepsilon_t^{-1} \circ T \circ \varepsilon_t$ which is independent of t.[9] The most

[9] Proof: $\varepsilon_t^{-1} T \varepsilon_t = \varepsilon_s^{-1} \Phi_{s,t} T \Phi_{t,s} \varepsilon_s = \varepsilon_s^{-1} \Phi_{s,t} \Phi_{t,s} T \varepsilon_s = \varepsilon_s^{-1} T \varepsilon_s$.

important symmetries of classical mechanics are those of Galilean spacetime, namely:

$$\begin{aligned}(q_i, p_i)_{1\leq i\leq N} \longrightarrow \qquad & (q_i + a, p_i) && \text{(Translation)}\\ & (Rq_i, Rp_i) && \text{(Rotation)}\\ & (q_i + ut, p_i + m_i u) && \text{(Galilei boost)}\end{aligned}$$

It is well-known that the Lebesgue or Liouville measure λ is invariant under these transformations, (which are just Euclidean transformations on phase space). Consequently (as is easy to check), the induced measure on $\mathcal{S}$ is invariant under the corresponding symmetry transformations on the solution manifold.

In Bohmian mechanics, the issue is a bit more subtle since the wave function itself transforms non-trivially under Galilean symmetries, namely as (Dürr and Teufel, 2009):

$$\begin{aligned}\Psi_t(q_1, ..., q_N) \longrightarrow \qquad & \Psi_t(q_1 - a, ..., q_N - a) && \text{(Translation)}\\ & \Psi_t(R^{-1}q_1, ..., R^{-1}q_N) && \text{(Rotation)}\\ & \mathrm{e}^{\frac{i}{\hbar}\sum_{i=1}^{N} m_i\left(uq_i - \frac{1}{2}u^2\right)}\Psi_t(q_1 - ut, ..., q_N - ut) && \text{(Galilei boost)}\end{aligned}$$

It is, however, evident that the $|\Psi|^2$-density is covariant under these transformations, guaranteeing the invariance of the typicality measure under Galilean symmetries.

In the upshot, the typicality measures in both classical mechanics and Bohmian mechanics are justified and tied to the dynamics by precise mathematical features: stationarity/equivariance and invariance/covariance under the fundamental spacetime symmetries. However, at least in classical mechanics, these conditions are not sufficient to determine the measure uniquely, or even rule out evidently inadequate choices such as a delta-measure concentrated on a stationary microstate. At the end of the day, part of what makes a measure compelling and allows it to play its normative role when it figures in typicality reasoning is that the choice doesn't seem biased or ad hoc or overly contrived. In other words, I do not believe that "soft" criteria can or should be completely avoided, and they aren't, in fact, in scientific practice. Attempts to axiomatize typicality measures (Werndl, 2013) seem misguided, not just because the particular proposals are uncompelling but because it is hard to see why any set of formal axioms should be more compelling than the very measures we use in physics.

Remark (Room for compromise)**.** We began this section by highlighting three questions on which the Mentaculus and the typically account disagree. We can now restate them more concretely:

i) Is the measure grounding the predictions of statistical mechanics a bona fide probability measure or a typicality measure?

ii) What is the epistemic and metaphysical status of the measure? Is it a theoretical postulate (a Humean law) on par with the dynamics, or does it formalize a way of reasoning about the laws?

iii) What expresses its epistemic or behavior-guiding function, the Principal Principle or Cournot's principle?

While I have defended a view that comes down on the opposite side of the Mentaculus on each of these points, I should emphasize that the answers to the three questions are logically independent (with the exception that a typicality measure essentially collapses PP into CP). For instance, it is possible to maintain that the (Humean) laws involve a typicality measure rather than a probability measure, that a probability measure expresses a way of reasoning (e.g., a principle of indifference) rather than an empirical postulate, or that the laws of nature include a bona fide probability measure in addition to deterministic laws, whose empirical and epistemic import comes from Cournot's principle. This leaves room for compromise, but also for misunderstandings since not everyone advocating for "typicality" or "'Humean chances" may have the same package deal in mind. And while there is always some value in compromise and moderation, the "extremal" positions are often the most interesting ones.

Chapter 6

The Logic of Typicality

Since typicality results are often formulated in terms of measure theory – or probability theory, i.e., *normalized* measures – the concepts of typicality and probability are easily conflated. This conflation usually happens by mistake, but can also come in the form of the criticism that "typical" is just another word for "very probable" despite attempts by some authors to make it into a bigger deal. In fact, as Wilhelm (2019) has already laid out in detail, typicality and probability are conceptually, formally, and metaphysically distinct. The goal of this section is to make the formal differences more explicit. For a more formal-logical discussion of typicality, see Crane and Wilhelm (2020).

6.1 Predicates, Propositions, and Extension Sets

When it comes to applications in physics, a fundamental premise of our discussion is that macroscopic facts are grounded in microscopic facts, with the latter being described by a suitable physical theory. We denote the microscopic state space of that theory by Ω and assume deterministic dynamics which determine a flow $\Phi_{s,t}$ on Ω such that $X(t) = \Phi_{s,t}(x)$ is the unique solution of the equations of motion with initial condition $X(s) = x$. In many cases, the micro-dynamics can even be described as a *dynamical system* $(\Omega, \Phi_{s,t}, \mu)$ with a (normalized) measure μ on the Borel sigma-algebra that is *stationary* under the flow $\Phi_{s,t}$.

The truth-value of any meaningful proposition φ about a physical system is determined by the microstate of the system or, ultimately, of the universe as a whole. We can thus associate φ with a predicate $P = P_\varphi$ on microstates whose extension

$$A_P = P_\varphi^{-1}(\{1\}) \tag{6.1}$$

is the subset of Ω which contains all and only those micro-configurations that realize P. In the terminology of statistical mechanics, P_φ is a macro-variable (with values in $\{0, 1\}$ corresponding to the truth values *false* and *true* respectively) and $A_P \subset \Omega$ the corresponding macro-region.[1] In other words, the kind of question we ask, in general,

[1] In principle, P may also refer to microscopic facts, in which case $A_P \subset \Omega$ will contain only few –

is not "what is the exact value of some continuous variable F" but "does the value of F lie in some range (y_1, y_2)?". The set Π of relevant predicates defines what we have called the *context* of our reasoning. It will always be closed under negation but not necessarily under logical conjunction and disjunction.

As we generalize this beyond classical mechanics, the implicit metaphysical assumption – which is rather an assumption about how the theoretical formalism connects to empirical facts – is that microscopic configurations "coarse-grain" to macroscopic facts, in the same sense in which, for instance, a moving configuration of Newtonian particles can coarse-grain to a planet circling the sun in an elliptical orbit. This is generally the case for *primitive ontology theories*, like Bohmian mechanics in the context of quantum physics. I believe that all serious candidates for a fundamental physical theory of the world should specify a primitive ontology or "local beables" in the sense of (Bell, 2004, Ch. 7), but this is a separate debate.

While a proposition $P_t(X)$ can be time-indexed – read: "the system in microstate X has the property P at time t – it is possible and convenient to evaluate all predicates at a common time $s = 0$ (usually the present or, if it exists, the beginning of time). To this end, we must simply note that:

$$X \in A_{P_t} \iff X(t) = \Phi_{0,t}(X) \in A_P \iff X \in \Phi_{0,t}^{-1}(A_P) \tag{6.2}$$

In other words, $\Phi_{0,t}^{-1}(A_P)$ is the set of micro-configurations at time $s = 0$ that realize P at time t. (Notably, both $t \leq 0$ and $t \geq 0$ are possible here). If we have a *stationary* measure μ on Ω, then

$$\mu(A) = \mu\left(\Phi_{0,t}^{-1}(A)\right) \tag{6.3}$$

for all t and all measurable sets A. This is another way to see why stationarity is such a critical – or at least very natural and convenient – requirement.

The basic intuition behind typicality is one of (very) large versus (very) small sets. To repeat: a property $P \in \Pi$ is typical within a reference set Ω if the great majority of elements in Ω instantiate P. To restate this within the framework just introduced: P is typical if the set A_P of microstates realizing P is very large. The exception set $A_P^c = \Omega \setminus \{x \in \Omega : \neg P(x)\}$ need not be empty but (in an appropriate sense) small or negligible. Symbolically:

$$\texttt{Typ}(P) \leftrightarrow \texttt{BIG}(A_P) \leftrightarrow \texttt{SMALL}(A_P^c) \tag{6.4}$$

Hence, formalizing "typical" in a given context does not require a whole range of numerical values but only a precise notion of "very large," respectively "negligibly small" sets. Measures in the sense of mathematical measure theory are but one – albeit a very natural and powerful – way to do so.

To flesh this out in more rigorous terms, I shall propose an axiomatization of *small*

or just one – microstate.

and *large* sets. I want to emphasize right away that this is not intended to be an exhaustive explication of typicality. I only claim that satisfying the axioms is a necessary but by no means sufficient condition for some logical structure to realize the notion of typicality. The exercise is somewhat akin to the abstract definition of a topology (let's say). It is illuminating and identifies minimal requirements for a set-structure to do the relevant job (supporting a sensible notion of continuity, convergence, etc.) but also admits of examples that are purely academic (e.g., the trivial topology in which anything converges to everything) and would never be accepted as a basis for physical or philosophical arguments. A more obvious analogy is that a sensible interpretation of probabilities need not regard all set-functions satisfying the Kolmogorov axioms as meaningful, or even metaphysically possible, examples. The deeper point is that typicality is not a purely logical concept but has a semantic and normative dimension that resists rigorous formalization.

6.2 A Theory of "Small Sets"

Let Ω be a base set, $|\Omega| \geq 2$, and $\Pi \subseteq \mathcal{P}(\Omega)$ a system of subsets that we want to evaluate as small, large, or neither. (Ω, Π) thus forms the *context* of a typicality reasoning. We now call $\mathcal{S} \subset \Pi$ a system of *small* sets if it satisfies the following axioms:

i) $\emptyset \in \mathcal{S}$

ii) $A \in \mathcal{S}, \Pi \ni B \subseteq A \Rightarrow B \in \mathcal{S}$

iii) $A, B \in \mathcal{S} \Rightarrow (A \cup B)^c \notin \mathcal{S}$

To make this more perspicuous, we introduce the two predicates on Π:

$$\begin{aligned} \texttt{SMALL}(A) &:\Leftrightarrow A \in \mathcal{S} \\ \texttt{BIG}(A) &:\Leftrightarrow A^c \in \mathcal{S}. \end{aligned}$$

That is, a set is BIG if and only if its complement is SMALL (but note that $\neg\texttt{SMALL}(A)$ does not imply $\texttt{BIG}(A)$). In terms of these predicates, the axioms read

i) $\texttt{SMALL}(\emptyset)$

ii) $\texttt{SMALL}(A), \Pi \ni B \subseteq A \Rightarrow \texttt{SMALL}(B)$

iii) $\texttt{SMALL}(A), \texttt{SMALL}(B) \Rightarrow \neg\texttt{BIG}(A \cup B)$

From these three axioms, we can immediately derive the following:

Lemma 6.2.1. *For all $A, B \in \Pi$:*

a) $\texttt{BIG}(\Omega)$

b) $\texttt{BIG}(A), A \subseteq B \Rightarrow \texttt{BIG}(B)$

c) $\mathtt{SMALL}(A), \mathtt{SMALL}(B) \Rightarrow \mathtt{SMALL}(A \cap B)$

d) $\mathtt{BIG}(A), \mathtt{BIG}(B) \Rightarrow \mathtt{BIG}(A \cup B)$

e) $\mathtt{BIG}(A), \mathtt{BIG}(B) \Rightarrow \neg\mathtt{SMALL}(A \cap B)$

f) $\mathtt{SMALL}(A) \Rightarrow \neg\mathtt{BIG}(A)$

Proof.

a) By i) since $\Omega = \emptyset^c$.

b) By ii) since $B^c \subseteq A^c$.

c) By ii) since $A \cap B \subseteq A$.

d) From c) since $(A \cup B)^c = A^c \cap B^c$

e) By iii), since $\mathtt{SMALL}(A^c), \mathtt{SMALL}(B^c) \Rightarrow \neg\mathtt{BIG}(A^c \cup B^c)$ and $A^c \cup B^c = (A \cap B)^c$, hence $\neg\mathtt{BIG}(A^c \cup B^c) \iff \neg\mathtt{SMALL}(A \cap B)$.

f) By i) and iii), since $\mathtt{SMALL}(A) \Rightarrow \neg\mathtt{BIG}(A \cup \emptyset = A)$.

□

In certain contexts, it makes sense to consider a stronger notion of "smallness," let's call it $\mathtt{SMALL}^*$, which is obtained by replacing axiom iii) with:

iii*) $\mathtt{SMALL}^*(A), \mathtt{SMALL}^*(B) \Rightarrow \mathtt{SMALL}^*(A \cup B)$.

Hence, while iii) only requires that the union of two small sets is not big, iii*) requires that it is still small. In technical terminology, the $\mathtt{SMALL}^*$-sets form an *ideal*, which corresponds to the notion of *negligible sets* sometimes found in the mathematical literature.

In some interesting cases, smallness (bigness) will even be closed under *countable* unions (intersections), that is

iii**) If $(A_i)_{i \geq 1}$ is a countable collection of $\mathtt{SMALL}^*$ sets, then $\mathtt{SMALL}^*\left(\bigcup\limits_{i=1}^{\infty} A_i\right)$.

The difference between finite and countable unions is important for technical purposes, but to keep things simple, we shall refrain from introducing another distinguishing notation.

Criteria for smallness

A trivial realization of $\mathtt{SMALL}^*$ (and *a forteori* $\mathtt{SMALL}$) is $\mathtt{SMALL}^*(A) \Leftrightarrow A = \emptyset$. This demonstrates consistency of the axioms but is otherwise uninteresting. Much more interesting and relevant ways to define smallness are the following:

1. **By Counting:** Ω a finite set with $|\Omega| = n$.

$$\texttt{SMALL}(A) :\Leftrightarrow |A| < k, \text{ for some fixed } k \leq \frac{n}{3}. \tag{6.5}$$

It is noteworthy that this most basic and intuitive definition satisfies only axioms i)-iii) but not iii*), i.e., smallness as defined by counting is not closed under unions.

2. **By Cardinalities:** Ω an infinite set.

$$\texttt{SMALL}^*(A) :\Leftrightarrow |A| < |\Omega|. \tag{6.6}$$

Examples: finite subsets of a countably infinite set (which satisfies axiom iii*)); countable subsets of an uncountably infinite set (which satisfies even iii**)); etc.

3. **Measure-theoretic:** $(\Omega, \mathcal{A}, \mu)$ a measure space with sigma-algebra $\mathcal{A} \supseteq \Pi$ and μ not trivial ($\mu(\Omega) \neq 0$).

Strong measure-theoretic notion:

$$\texttt{SMALL}^*(A) \Leftrightarrow \mu(A) = 0. \tag{6.7}$$

This satisfies axiom iv’), i.e., closedness under countable unions. $\texttt{BIG}^*$ sets in this sense contain *almost all* elements of Ω.

Weak measure-theoretic notion:

$$\texttt{SMALL}(A) :\Leftrightarrow \frac{\mu(A)}{\mu(\Omega)} < \epsilon, \text{ for some fixed } \epsilon \leq \frac{1}{3}. \tag{6.8}$$

$\texttt{BIG}$ sets in this sense contain *nearly all* elements of Ω.

Family of measures: $\mathcal{M}$ a set of measures on $(\Omega, \mathcal{A})$.

$$\texttt{SMALL}(A) :\Leftrightarrow \sup_{\mu \in \mathcal{M}} \frac{\mu(A)}{\mu(\Omega)} \leq \epsilon. \tag{6.9}$$

In other words, $\texttt{SMALL}(A)$ iff $\mu(A) \leq \epsilon$ for *all* measures μ in $\mathcal{M}$.

Non-normalizable measures: The previous two definitions include the case $\mu(\Omega) = \infty$ with the convention $\frac{1}{\infty} = 0$ and $\frac{\infty}{\infty} = 1$. Hence, (6.8) becomes

$$\texttt{SMALL}^*(A) : \Longleftrightarrow \ \mu(A) < \infty = \mu(\Omega), \tag{6.10}$$

which is closed under finite unions.

4. **Dimensional:** Ω a normal topological space.

$$\texttt{SMALL}^*(A) :\Leftrightarrow \dim(\bar{A}) < \dim(\Omega) \tag{6.11}$$

where dim is the topological dimension and $\bar{A}$ denotes the closure of A.

5. **Topological:** Ω a connected T_1 topological space.[2]

$$\texttt{SMALL}^*(A) :\Leftrightarrow A \text{ is a } \textit{nowhere dense} \text{ set.} \tag{6.12}$$

A set A is *nowhere dense* if its closure has empty interior, or, in other words, if for any neighborhood $U \subseteq \Omega$, there exists a non-empty open set $V \subseteq U$ such that $A \cap V = \emptyset$.

Alternatively: Ω a *Baire space.*[3]

$$\texttt{SMALL}^*(A) :\Leftrightarrow A \text{ is a } \textit{meagre} \text{ set.} \tag{6.13}$$

A set is called *meagre* if it is a countable union of nowhere dense sets. Meagre sets are also called sets of first *Baire category.* The complement of a meagre set is sometimes called *comeagre.*

The crucial difference between (6.12) and (6.13) is that “meagreness” is closed under countable unions (meagre sets form a sigma-ideal) while nowhere denseness is only closed under finite unions. $\mathbb{Q} \subset \mathbb{R}$ is an example of a set that is meagre but dense (and hence not nowhere dense).

Some remarks on the interrelations between these definitions:

- For countable (including finite) Ω, the counting-theoretic definition corresponds to the weak measure-theoretic definition (6.8) with the counting measure.
- If Ω is an uncountable set and the measure μ has no discrete part (i.e., $\mu(\{x\}) = 0, \forall x \in \Omega$), the cardinality definition is stronger than the measure-theoretic one: All countable sets have measure zero (while, in general, not all measure zero sets are countable).
- If $\Omega \cong \mathbb{R}^n$, all measurable subsets of dimension $< n$ have Lebesgue measure zero.

The topological and measure-theoretic notions are, in general, orthogonal. In particular, there exist not only dense Lebesgue null sets (e.g., $\mathbb{Q} \subset \mathbb{R}$) but also nowhere dense subsets of the unit interval with measure arbitrarily close to 1 (“fat Cantor sets”).

Example (Smith-Volterra-Cantor set)**.** The best known example of a nowhere dense set with positive measure is the *Smith-Volterra-Cantor set.* It is constructed as follows:

[2] The T_1 separation axiom states that for all $x \neq y$ there exists a neighborhood U of y such that $x \notin U$. Together with Ω connected, this ensures that one-element subsets are nowhere dense.

[3] A Baire space is a topological space in which the union of every countable collection of closed sets with empty interior has empty interior. This ensures that Ω is not meagre itself.

In the first step, remove from the unit interval $[0,1]$ the middle open interval of length $1/4$, leaving $S_1 = [0,3/8] \cup [5/8,1]$.

In the n'th step, remove from each of the remaining 2^{n-1} intervals the middle open interval of length $\left(\frac{1}{4}\right)^n$, leaving a union S_n of 2^n closed connected intervals.

The Smith-Volterra-Cantor set is then $S_\infty = \bigcap\limits_{n=1}^{\infty} S_n$.

As a countable intersection of closed sets, S_∞ is itself a closed set. And it contains no open interval (every interval is broken up at some step of the iterative process), hence it has empty interior, hence it is nowhere dense. However, we have removed in total a set of measure $\sum\limits_{n=1}^{\infty} \frac{2^{n-1}}{2^{2n}} = \sum\limits_{n=1}^{\infty} \frac{1}{2^{n+1}} = \frac{1}{2}$ from the unit interval, so that $\lambda(S_\infty) = 1 - \frac{1}{2} = \frac{1}{2}$.

The incommensurability of topological and measure-theoretic criteria poses a challenge to my claim that typicality is one unified concept. Sets that are small in the cardinality-theoretic or dimensional sense are also small with respect to natural measures. In this sense, cardinality and dimension provide stronger criteria for typicality. The nowhere-denseness of a set, however, bares no relation to its content (except that a nowhere dense set cannot have *full* measure if the measure is strictly positive, i.e., if $\mu(U) > 0$ for any non-empty open set).

My view is that the topological notions introduced above are indeed *not* related to typicality in the sense discussed in this thesis. Topology is, roughly speaking, about closeness and separation of points, not about their quantity. A nowhere dense set A need not be small in the sense of containing few elements. Rather its points don't accumulate in Ω: Every open neighborhood of Ω contains points that are topologically separated from A (i.e., not in A or on its boundary).

If we think of physical properties and the microstates[4] realizing them, there is certainly an idea of counterfactual robustness associated with the basic topological notion of open sets, viz. that the property is robust under small perturbations. We may say that a microstate x is "good" if not only x itself realizes P but all points that are sufficiently close to x, as well. If the set of microstates realizing P is nowhere dense, it implies that there are no good points. If its complement is nowhere dense, it means that good points are spread out all over Ω. But this notion of "goodness" is doing both too much and too little for the purposes of a typicality reasoning. Too little because open sets can be arbitrarily small. And too much because the "bad" configurations not realizing P can be very few and special ones, even if they are "all over the place" (like rational points in the continuum) or concentrated in some (small) region of phase space.

[4] "Microstate" here need not refer to particle configurations. In cosmology, for instance, it may be a metric realizing certain geometric properties of spacetime.

In the end, the touchstone is whether the topological definitions can ground the relevant normative implications of typicality facts. If a physical phenomenon is realized for all but a meagre/nowhere dense set of initial conditions, I might be satisfied that it doesn't require a suspicious fine-tuning, but wouldn't go as far as to consider it conclusively explained. At least not to the extent that any further questions seem void, as I take to be the case for typical phenomena.

With all that said, I cannot completely dispel the charge that my arguments are begging the question by presupposing measure-theoretic intuitions. But I would insist, again, on a semantic aspect of both typicality and mathematics. The meaning of a formal concept matters for its ability to express typicality.

Conditional Measures

What we don't quite get from the abstract axioms but the concrete realizations of' 'smallness" discussed above is a notion of *conditional typicality*. The most interesting and useful one is given in terms of conditional measures. If $\mu(A) > 0$, then the conditional measure $\mu(\cdot \mid A)$ is defined by

$$\mu(B \mid A) = \frac{\mu(A \cap B)}{\mu(A)} \text{ for } B \in \mathcal{A} \tag{6.14}$$

In "benign" cases, it is also possible to conditionalize a measure on a null-set, but for simplicity, we shall generally assume $\mu(A) > 0$.

Many relevant applications of typicality refer, in fact, to conditional typicality. An event B may not be typical/atypical *simpliciter* but become typical/atypical *given* A. Conditionalizing on some event/macrostate may even make an atypical event typical or vice versa. For instance, entropy-increase in a closed system is typical given a low-entropy initial state, but it is atypical simpliciter since nearly all possible microstates are already in a state of maximal entropy. Moreover, from a fundamental physical point of view, many predicates – e.g., "the relative frequency of *heads* in series of $N = 1000$ coin tosses is approximately 1/2"– do not even make sense without pertinent boundary conditions, because most possible universes may not contain any coins, to begin with. (And some may contain coins, but fewer than 1000 tosses are ever made.) In such cases, the relevant boundary conditions are often left implicit.

6.3 Typicality Measures

The following proposition specifies the condition under which a system of small sets can be characterized by the measure-theoretic notion of null sets.

Proposition 6.3.1. *Let $\sigma(\mathcal{S})$ be the sigma-algebra generated by $\mathcal{S}$ (i.e., the smallest sigma-algebra containing all small sets). Then there exists a measure ν on $\sigma(\mathcal{S})$ such that*

$$\texttt{SMALL}(A) \Leftrightarrow \nu(A) = 0 \tag{6.15}$$

if and only if $\texttt{SMALL}(\cdot)$ *is closed under countable unions.*

Proof. For any measure ν, the null-sets form a sigma-ideal closed under countable unions. This follows from the σ-subadditivity of measures: If $(A_i)_i \geq 1$ is a countable family with $\nu(A_i) = 0 \ \forall i \geq 1$, then $\nu(\bigcup_{i\geq 1} A_i) \leq \sum\limits_{i=1}^{\infty} \nu(A_i) = 0$. Conversly, suppose that $\texttt{SMALL}(\cdot)$ is closed under countable unions. Then the set-function

$$\tilde{\nu}(A) = \begin{cases} 0; & \text{if } \texttt{SMALL}(A) \\ 1; & \text{if } \texttt{BIG}(A) \end{cases}$$

extends to a (normalized) measure ν on $\sigma(\mathcal{S})$. □

Note that the measure obtained above is not quite what we want, since it may not be defined on all sets in Π. It is, however, sufficient to identify all small and big sets (while all $A \in \Pi \setminus \sigma(\mathcal{S})$ are neither).

The following negative result holds for the weaker measure-theoretic notion of smallness:

Definition 6.3.2. A measure μ on $(\Omega, \mathcal{A})$ is called *non-atomic* (and $(\Omega, \mathcal{A}, \mu)$ is called a non-atomic measure space) if

$$\mu(A) > 0 \Rightarrow \exists B \subset A : 0 < \mu(B) < \mu(A).$$

In this case, a kind of intermediate-value theorem holds for the measure: by *Sierpinski's theorem*, there exists for any $A \in \mathcal{A}$ and $p \in [0, \mu(A)]$ a $B \in \mathcal{A}$ with $\mu(B) = p$.

Proposition 6.3.3. *Let* $(\Omega, \mathcal{A}, \mu)$ *a nonatomic measure space. Let* $\epsilon \leq \frac{1}{2}\mu(\Omega)$ *and define*

$$\texttt{SMALL}(A) :\Leftrightarrow \mu(A) < \epsilon, \ A \in \mathcal{A} \tag{6.16}$$

Then there exists $A, B \in \sigma$ *with* $\texttt{SMALL}(A), \texttt{SMALL}(B)$ *but* $\neg\texttt{SMALL}(A \cup B)$*. Hence,* $\texttt{SMALL}$ *cannot be closed under unions if defined on the entire sigma-algebra.*

Proof. Since $\mu(\Omega) \geq 2\epsilon$, there exists (by Sierpinski's theorem) a measurable B with $\mu(B) = \frac{3}{4}\epsilon$. And since $\mu(B^c) \geq \frac{5}{4}\epsilon$ there exists $A \in \mathcal{B}^c$ with $\mu(A) = \frac{3}{4}\epsilon$. Hence $\mu(A) = \mu(B) < \epsilon$ but since A and B are disjoint, $\mu(A \cup B) = \mu(A) + \mu(B) = \frac{6}{4}\epsilon > \epsilon$. □

This result is the reason why not all notions of "smallness," and hence typicality, can be reduced to the measure-theoretic one. Wilhelm (2019) provides the following instructive example: Consider $\Omega = \mathbb{N}$ with the cardinality-theoretic criterion of smallness, i.e., $\texttt{SMALL}^*(A) \iff |A| < \infty$. For $n \geq 1$, let $A_n = \{1, \ldots, n\}$. Then $(A_n)_{n\in\mathbb{N}}$ is an ascending sequence of finite (and thus small) sets with $\bigcup\limits_{n=1}^{\infty} A_n = \mathbb{N}$. Now suppose there was a measure μ with $\mu(A_n) < \epsilon < \mu(\mathbb{N})$ for all n. By the upwards continuity of measures, we would have $\mu(\mathbb{N}) = \mu\left(\bigcup\limits_{n=1}^{\infty} A_n\right) = \lim\limits_{n\to\infty} \mu(A_n) \leq \epsilon$ and hence a contradiction.

Nonetheless, as discussed in Section 3.1, it is usually possible to reconcile (6.8) (the weak measure-theoretic criterion for smallness) with axiom iii*) (closedness under unions) by restricting our considerations to an appropriate class Π of relevant sets. This is to say that we don't define the smallness predicate on the entire sigma-algebra of measurable sets, but care only about a more limited class of subsets of Ω. The following proposition states that closedness under unions is then equivalent to the existence of a strongest typicality fact, i.e., a smallest `BIG` set $B_\Omega \in \Pi$. We get, moreover, a notion of (epistemic) "resilience" or "stability" that is central to Leitgeb's *Stability Theory of Belief*: No fact compatible with B_Ω can make B_Ω atypical.

Proposition 6.3.4. *(Leitgeb, 2017, Thm. 7, p. 121) Suppose Π is finite/countable and*

$$\texttt{BIG}(A) \iff \mathbb{P}(A) > 1-\epsilon,\ A \in \Pi$$

for a normalized measure $\mathbb{P}$. Then the following two conditions are equivalent:

i) $\texttt{BIG}(\cdot)$ is closed under finite/countable intersections.

ii) There exists a smallest `BIG` set B_Ω such that

$$\texttt{BIG}(A) \iff B_\Omega \subseteq A. \tag{6.17}$$

Evidentally, B_Ω is the union of all `BIG` sets. And it is also $\mathbb{P}$-stable*, meaning that for any A with $\mathbb{P}(A) > 0$ and $\emptyset \neq B_\Omega \cap A \in \Pi$:*

$$\mathbb{P}(B_\Omega \mid A) > \frac{1}{2}. \tag{6.18}$$

Proof. Assuming i), $B_\Omega := \bigcap_{\texttt{BIG}(A)} A$ is the desired set satisfying ii).
Assuming ii), let $(A_i)_{i\geq 1}$ be a countable collection of `BIG` sets. Then, $\forall i \geq 1 : \texttt{BIG}(A_i) \iff \forall i \geq 1 : B_\Omega \subseteq A_i \iff B_\Omega \subseteq \bigcap_i A_i \iff \texttt{BIG}(\bigcap_i A_i)$. Hence, $\texttt{BIG}(\cdot)$ is closed under intersections.

To prove (6.18), we consider for $\mathbb{P}(A) > 0$ the conditional probability

$$\mathbb{P}(B \mid A) = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(A)} = \frac{\mathbb{P}(B \cap A)}{\mathbb{P}(B \cap A) + \mathbb{P}(B^c \cap A)} = \left(1 + \frac{\mathbb{P}(B^c \cap A)}{\mathbb{P}(B \cap A)}\right)^{-1}.$$

Now we note the following:

Lemma 6.3.5. *If $A \in \Pi$ with $A \subset B_\Omega$, then $\mu(A) > \epsilon$.*

For else, $\mathbb{P}(A^c) \geq 1 - \epsilon \Rightarrow \texttt{BIG}(A^c)$, but $B_\Omega \nsubseteq A^c$ in contradiction to ii). Hence, if $\emptyset \neq B_\Omega \cap A \in \Pi$, we have $B_\Omega \cap A \subset B_\Omega \Rightarrow \mathbb{P}(B_\Omega \cap A) > \epsilon$, while $B_\Omega^c \cap A \subset B_\Omega^c \Rightarrow \mathbb{P}(B_\Omega^c \cap A) \leq \epsilon$, and thus $\mathbb{P}(B_\Omega \mid A) > \frac{1}{2}$.

□

Equivalence of measures

Absolute Continuity

Let $(\Omega, \mathcal{A})$ a measurable space. A measure ν is called *absolutely continuous* with respect to another measure μ (notation: $\nu \ll \mu$) if

$$\mu(A) = 0 \Rightarrow \nu(A) = 0, \ \forall A \in \mathcal{A}. \tag{6.19}$$

If $\nu \ll \mu$ and $\mu \ll \nu$, then the two measures are equivalent in that they distinguish the same null sets – and thus the same $\texttt{SMALL}^*$ sets in the sense of (6.7).

If $\Omega = \mathbb{R}^n$ (with the Borel sigma-algebra), "absolute continuity" is implicitely understood relative to the Lebesgue measure λ unless μ is otherwise specified. Then $\nu \ll \lambda$ implies: For all $\epsilon > 0$ there exist $\delta > 0$ such that $\lambda(A) < \delta \Rightarrow \nu(A) < \epsilon$. Informally: if sets are sufficiently small with respect to the Lebesgue measure, they will also be small with respect to the absolutely continuous measure ν.

In general, if $\nu \ll \mu$, there exists an integrable function g (unique up to sets of measure zero) such that

$$\nu(A) = \int_A g \, \mathrm{d}\mu, \ \forall A \in \mathcal{A} \tag{6.20}$$

In other words, ν has a *density* with respect to μ. This is the *Radon-Nikodym theorem* and the density g is also called the *Radon-Nikodym derivative* of ν w.r.t. μ.

Total variation of measures and typicality thresholds

Absolute continuity provides an equivalence of typicality measures in the sense of the strong measure-theoretic criterion (6.7). We now want to clarify in what sense different measures can provide an equivalent notion of smallness in the sense of the weaker criterion (6.8).

There are two ways to think about this equivalence. If the conditions from Proposition 6.3.4 apply, smallness on Π is already determined by a smallest $\texttt{BIG}$ set, respectively a largest $\texttt{SMALL}$ set, i.e., by the logical/set-theoretic structure of $(\Pi, \mathcal{S} = \{A \subseteq B_\Omega^c\})$. In terms of a (normalized) measure μ, the (largest possible) threshold value for smallness is then evidently

$$\texttt{SMALL}(A) \iff \mu(A) < \epsilon := 1 - \mu(B_\Omega), \text{ for } A \in \Pi.$$

But if this ϵ is reasonably small, i.e., $\mu(B_\Omega) \approx 1$, we will also have $\tilde{\mu}(B_\Omega) \approx 1$ for any measure $\tilde{\mu}$ that doesn't deviate too radically from μ.

In practice, however, we usually find ourselves in the opposite situation that we don't know B_Ω (if it exists) but rely on a natural measure to determine what is typical and atypical. And we have said that it doesn't seem reasonable, in general, to specify a sharp threshold value above which sets are no longer considered as small. Instead, we should understand smallness – and thus typicality – as a *vague* predicate (which is better captured by the more informal condition $\mathtt{SMALL}(A) \iff \mu(A) \approx 0$).

What we can do – regardless of the applicability of (6.18) and more in line with practical applications – is to specify some $\epsilon > 0$ such that $\mu(A) < \epsilon$ is a *sufficient* condition for typicality in the given context, i.e.,

$$\mu(A) < \epsilon \Rightarrow \mathtt{SMALL}(A),\ A \in \Pi. \tag{6.21}$$

Similarly, we could specify $\epsilon < \Upsilon < 1 - \epsilon$ such that $\mu(A) < \Upsilon$ is a *necessary* condition, i.e., $\mathtt{SMALL}(A) \to \mu(A) < \Upsilon$, while remaining agnostic about the range $\mu(A) \in [\epsilon, \Upsilon]$.

Remark. For proper typicality arguments, we will always require $\epsilon \ll 1$, but in different contexts, the relevant orders of magnitude may be different. Indeed, in typicality results, the pertinent estimates are generally about *orders of magnitude* (e.g., in terms of powers of N in LLN results) and we rarely, if ever, find ourselves quibbling over exact numerical values, e.g., whether a measure of 10^{-24} is small when referring to possible micro-configurations of a gas or whether we should insist on values below 0.9975×10^{-24}.

If we are primarily interested in identifying small sets in the sense of (6.21), then, again, many different measures could do the job. Many measures will agree on the smallness of a set, provided that they do not differ too radically from one another. Let us try to make this more precise.

A convenient metric on the space of normalized measures on $(\Omega, \mathcal{A})$ is the *total variation distance*

$$d_{TV}(\mu, \nu) = \sup_{A \in \mathcal{A}} |\mu(A) - \nu(A)|. \tag{6.22}$$

Simply put, this is the maximal disagreement between the values that two measures can assign to the same set. Evidently, if $d_{TV}(\mu, \nu) = \delta$, then

$$\nu(A) < \epsilon \Rightarrow \mu(A) < \epsilon + \delta \tag{6.23}$$

Hence, if $\mu(A) < \epsilon + \delta$ is sufficient for $\mathtt{SMALL}(A)$, then $\nu(A) < \epsilon$ is sufficient, as well.

More generally: Let $\mathcal{M}$ be a set of normalized measures with $\mathrm{diam}_{TV}(\mathcal{M}) = \sup\{d_{TV}(\mu, \nu) : \mu, \nu \in \mathcal{M}\} = \delta$.

$$\begin{aligned} &\text{If } \sup_{\mu \in \mathcal{M}} \mu(A) < \epsilon + \delta \Rightarrow \mathtt{SMALL}(A) \\ &\text{Then } \nu(A) < \epsilon \Rightarrow \mathtt{SMALL}(A) \text{ for any } \nu \in \mathcal{M}. \end{aligned} \tag{6.24}$$

Similarly, if $\inf_{\mu \in \mathcal{M}} \mu(A) > \Upsilon \Rightarrow \neg\mathtt{SMALL}(A)$, then $\nu(A) > \Upsilon + \delta \Rightarrow \neg\mathtt{SMALL}(A)$. Hence, every representative of the class of measures can provide necessary and sufficient criteria for typicality.

A bound in total variation is actually too strong. Given a reference measure μ, all we need is a bound on

$$\sup\left\{|\nu(A)-\mu(A)| : \mu(A)<\epsilon\right\}. \tag{6.25}$$

If $\nu \ll \mu$ with a density (Radon-Nikodym derivative) g, then (6.25) can be written as $\sup\left\{|\int \mathbb{1}_A(g-1)\mathrm{d}\mu| : \mu(A)<\epsilon\right\}$ (evidentally, μ has the constant density 1 with respect to itself). With the *Hölder inequality*, we thus get for all $p\in[1,\infty]$:

$$\sup_{\mu(A)<\epsilon}|\nu(A)-\mu(A)| = \sup_{\mu(A)<\epsilon}\left|\int \mathbb{1}_A(g-1)\,\mathrm{d}\mu\right| \leq \epsilon^{1-\frac{1}{p}}\|g-1\|_p \tag{6.26}$$

On the right-hand-side, we have the so-called L^p-norms defined as $\|f\|_p = \left(\int|f|^p\mathrm{d}\mu\right)^{1/p}$ for $p<\infty$, and $\|f\|_\infty = \operatorname{ess\,sup}_{x\in\Omega}|f(x)|$ (the *essential* supremum is the least upper bound on $|f(x)|$ *almost everywhere*, i.e., with the possible exception of μ-null sets).

The point of this mathematical exercise is that, since ϵ is very small, g would have to differ radically from the constant density – in the sense of these L^p-norms – to assign a large measure to μ-small sets. (However, $\|g-1\|_p$ could be infinite for high values of p, so the higher the regularity of g, the stronger the bound from (6.26).)

Example. Consider the model of an ideal gas of $N\approx 10^{24}$ particles in a box. Let A be the macro-region for which the left-hand side of the box contains significantly more particles than the right-hand-side. Then, with respect to the Lebesgue measure $\lambda(A)\approx 10^{-10^{24}}$. However, a measure of less than $10^{-10^{23}}$ would certainly be sufficient for $\texttt{SMALL}(A)$. Thus, if ν is absolutely continuous with bounded density g, this density would have to peak at values of $g(x)\gtrsim 10^{(10^{24}-10^{23})}\approx 10^{10^{24}}$ to disagree on the smallness of that set.

Chapter 7

Other Applications of Typicality

In this short chapter, we will discuss applications of typicality beyond statistical mechanics and probability theory. On the one hand, this will emphasize the wide scope and philosophical potential of typicality. On the other hand, in the following examples, the appeal to probabilistic concepts is highly questionable. The examples will thus help to further clarify the difference between typicality and probability.

7.1 Typicality and Well-Posedness

Deterministic laws are, in general, given as (differential) equations of motion that lend themselves to a *well-posed* initial value problem. The terminology goes back to Hadamard (1902), who required of *un problème bien posé* that (i) a solution exists, (ii) the solution is unique, and (iii) that the solutions depends continuously on the initial data.

At least the first two conditions are certainly necessary for determinism, i.e., for the history of a closed physical system – on the fundamental level the universe – to be fully determined given its complete physical state at one moment in time. A famous example for the failure of condition ii) in the framework of classical mechanics is *Norton's dome* (Norton, 2008). Technically, what happens in this case is that the Hamiltonian vector field fails to be Lipschitz-continuous, thus violating a premise of the Pascal-Lindelöff theorem establishing existence and uniqueness of solutions of ordinary differential equations. *Existence* of solutions as required by Hadamard's condition (i) means, in general, *global* existence for all times. A solution $X(t)$ that cannot be extended beyond a bounded (or half-bounded) interval (t_0, t_1), $t_0 > -\infty \vee t_1 < +\infty$ indicates the formation of a *singularity* at which the equations of motion break down.

In Newtonian gravity, this happens when two point particles collide so that the gravitational force between them diverges: $\lim\limits_{r \to 0} \frac{Gm_1 m_2}{r^2} = +\infty$. However, it is known that initial conditions leading to such collision singularities are atypical; they form a set of Lebesgue measure zero that is also topologically meagre (Saari, 1971a, 1973).

More surprising is the possibility of non-collision singularities in the N-body problem, where particles go off to infinity in finite time. (The time-reversal of such solutions – corresponding to particles suddenly appearing in space – has also been discussed as an example of Newtonian indeterminism, see, e.g., Earman (1986, Ch. 3)). Solutions with non-collision singularities are known to exist for $N \geq 5$ (Xia, 1992), but not for $N \leq 3$ (Painlevé, 1897). The case $N = 4$ is an open mathematical problem. Somewhat ironically, this is the only N for which we have a rigorous proof that initial conditions leading to non-collision singularities (if they exist at all) must also form a Lebesgue null set. Saari conjectures that this holds true for all $N \geq 4$ (Saari, 2005, p. 221) and intuitively, it seems clear that only very conspiratorial behavior could lead to particles being accelerated to infinity in finite time.

For the guiding equation of Bohmian mechanics, the solution theory is more settled. For sufficiently "nice" wave functions Ψ, singularities can occur only if the particle configuration runs into a node of the wave function where the velocity field diverges. However, equivariance of the $|\Psi|^2$-measure under the Bohmian flow already implies that Bohmian trajectories tend to avoid regions where $\Psi \approx 0$. And indeed, global existence of solutions has been proven for almost all initial conditions (i.e., a set of measure 1) relative to this natural typicality measure (Berndl et al., 1995; Teufel and Tumulka, 2005).

So we note that actual scientific practice aims at establishing existence and uniqueness of solutions for almost all, i.e., typical initial conditions[1] when singularities are mathematically possible. Such results are generally regarded as satisfactory, establishing that the laws are sound and deterministic – despite pathological "counterexamples" that receive more attention in the philosophical than in the physical literature.

There is certainly an empirical rationale here. On the one hand, atypical micro-configurations leading to singularity formation would be virtually impossible to create *in practice*. On the other hand, no empirical evidence could ever justify a belief that the actual microstate of the universe is one of these "bad" configurations. However, if we want to take the laws seriously as candidates for fundamental laws of nature, subjective belief and practical impossibility are not the right concepts, and one would rather make the claim that singular solutions are *physically impossible*. A singularity, after all, does not mean that the evolution of the universe suddenly stops but that the laws themselves are breaking down.

If we take this position, we make a typicality reasoning not with respect to a reference class of nomologically possible worlds but with respect to a mathematical solution space. And we are satisfied with the mathematical expressions of the law if the relation between formal solutions and possible worlds is not quite one-to-one provided that the solutions we regard as unphysical are also atypical. Probability, I claim, does not support an analogous reasoning. Impossible initial conditions are not unlikely but impossible. And whether a particular solution runs into a singularity is not random but a mathematical fact.

[1] In the strong sense of "all initial conditions except for a set of measure zero."

The situation as described for Newtonian gravity and Bohmian mechanics should be compared with general relativity (GR), where singularity theorems establish the existence of spacetime singularities (in the form of geodesic incompleteness) under very general conditions (see, e.g., Hawking and Ellis (1973)). In other words, singularity formation in GR seems to be generic rather than atypical. Although some of these singularity theorems are actually "predictive" – in that they are taken to establish the inevitability of a Big Bang – they must be considered as negative results from a foundational perspective, pointing to an intrinsic limitation of GR and the necessity of a more fundamental theory of spacetime.

7.2 Typicality and Fine-Tuning

We can say: a feature of our universe that is atypical in the sense discussed so far requires a *fine-tuning* of initial micro-conditions. The initial conditions of the universe would have had to be extremely special ones in order to produce it. The feature could be a (statistical) regularity instantiated in our universe, in which case there is little debate, at least among physicists, that fine-tuning is bad – if only for the reason that a theory could be fine-tuned to account for virtually anything. In physics, however, the term fine-tuning is more commonly used when referring to the singular value of some physical quantity that we observe in our universe, including parameters that have not been "produced" by dynamical evolution but have the status of a constant of nature in the theory. This raises further questions about the applicability of a typicality reasoning.

The flatness problem

A famous example of a fine-tuning problem is the *flatness problem* in standard Big Bang cosmology (which is said to have been resolved by inflationary cosmology). The "puzzle," in a nutshell, is that the mass/energy density of our universe (ρ) is very close to the "critical value" (ρ_c) required for a flat spatial geometry on cosmological scales. Moreover, the energy density departs very quickly from the critical value as the universe expands. The ratio ρ/ρ_c is commonly denoted by Ω, and while $\Omega \approx 1$ today, the deviation from unity would have had to be even $\approx 10^{60}$ times smaller at the Planck time, shortly after the Big Bang. More precisely, from the Friedmann equation,

$$(\Omega^{-1} - 1)\rho a^2 = -\frac{-3kc^2}{8\pi G}, \tag{7.1}$$

where a is the so-called scale factor and $k \in \{\pm 1, 0\}$ indicating negative, positive, or flat curvature, respectively. However, $\rho \propto a^{-3}$ for matter and $\propto a^{-4}$ for radiation, so if $\rho a^2 \propto a^{-1}$ has decreased by a factor of 10^{60} as the universe expanded, $(\Omega^{-1} - 1)$ must have increased accordingly (the right-hand-side of the equation is constant).

It is questionable whether this is really puzzling or rather what we called a Morgenbesser case in Ch. 1.2: The value of Ω had to be *something*, after all, and while 1 may seem special to us, it is as good (or rather as atypical) as any other. Moreover, several authors have pointed out that since Ω always tends to unity as we approach the big bang ($a \to 0$), there is always a time in the very early universe at which it would look "fine-tuned" (Coles and Ellis (1997, p. 22); Lake (2005); Helbig (2012)). After all, in classical (i.e., non-quantum) cosmology, there is nothing special about the Planck time that many statements of the flatness problem take as reference point. On the other hand, one could argue that we are currently experiencing a very special period in the history of our universe in which $\Omega \approx 1$, but this turns the flatness problem into a problem of self-location rather than fine-tuning, requiring somewhat different considerations (including anthropic ones) that go beyond the scope of this particular discussion.

In any case, the interesting philosophical debate is over which features of our universe are a valid target of scientific explanation versus acceptably brute. The flatness of our universe *per se* may not be a good explanandum. It seems, however, that if the value of Ω in the early universe had been slightly different from what it was, the universe would have either recollapsed too quickly or expanded too fast to allow for the formation of stars and galaxies (Lewis and Barnes, 2016, pp. 164-167). The relevant phenomenon that the standard big bang theory fails to explain is thus not the "special" numerical value of Ω, but the abundance of stars and galaxies in our universe. This is what the energy density appears to be fine-tuned for.

That said, a proper typicality analysis should consider the possible "initial" configurations of the matter and metric fields, not the possible values of Ω in some abstract parameter space. Unfortunately, whenever we are dealing with a field theory (such as general relativity) the fundamental state space Γ is infinite-dimensional, which makes the construction of a natural typicality measure difficult (see Curiel (2015) for a good philosophical discussion).

There exists, however, a canonical measure (the *GHS* measure) on the *reduced* phase space ("minisuperspace") of the Friedmann–Lemaître–Robertson–Walker models used in standard big bang cosmology. And with respect to this measure, a flat universe turns out to be, in fact, *typical*:

> "Thus for arbitrarily large expansions (and long times), and for arbitrarily low values of the energy density, the canonical measure implies that almost all solutions of the Friedmann-Robertson-Walker scalar equations have negligible spatial curvature and hence behave as $k = 0$ models." (Hawking and Page, 1988, pp. 803-4)

> "[T]he measure is entirely concentrated on exactly flat universes; universes with nonvanishing spatial curvature are a set of measure zero. [...] Therefore, our interpretation is clear: almost all universes are spatially flat." (Carroll and Tam, 2010, p. 18)

The second statement, in particular, may be a little too strong. In a recent critical discussion, McCoy (2017) points out that without a regularization which he finds questionable, the result is rather that for any value κ_* of the curvature parameter $\kappa = \frac{k}{a^2}$, FLRW spacetimes with $\kappa < \kappa_*$ form a set of infinite measure, while those with $\kappa \geq \kappa_*$ form a set of finite measure. This is a perfectly valid standard of typicality as we discussed in 6. In particular, a typicality measure – in contrast to a probability measure – does not have to be normalizable.

Of course, in the present case, we should then say that "nearly flat" (rather than "exactly flat") spacetimes are typical. McCoy objects, however, that the threshold value κ_* for "nearly flat" is arbitrary – that any observed curvature could thus be deemed either typical or not small enough. I don't see this problem. On the one hand, if we are asking "Why does the large scale structure of our universe look so flat?", the explanandum itself is vague. On the other hand, I have argued that the relevant phenomenon to be explained is actually structure formation, which makes the upper bound κ_* not arbitrary. The universe must have been *flat enough* to allow for structure formation, and according to the GHS measure, this is typical.

McCoy concludes that mainly due its non-normalizability, the GHS measure "cannot be used to make typicality arguments in this context" (McCoy, 2017, p. 1251). I disagree. However, physicists with great expertise have also expressed skepticism about the justification for using this particular typicality measure (see, e.g., Schiffrin and Wald (2012)).

While typicality helps to avoid the conceptual problems associated with probabilistic reasoning in cosmology, technical challenges remain. Those are mitigated by the fact that we *do not* have to insist on normalized measures and thus regularization procedures that tend to cause most of the ambiguities. Nonetheless, for a particle ontology (as we assume throughout most of our discussions), the mathematical side is usually much simpler.

Fine-Tuning of the natural constants

Another famous fine-tuning problem is the fine-tuning of the natural constants. For instance, if the strength of the strong nuclear force compared to that of electromagnetism would have been significantly different, heavy elements couldn't have formed in stellar fusion processes. Either most of the hydrogen would have been burned in the very early universe or stellar nucleosynthesis would have been much less efficient (see, e.g., Barrow and Tipler (1986); Lewis and Barnes (2016)).

Notably, the explanandum here is not that the value of the fine structure constant is close $\frac{1}{137}$. To wonder about something like that strikes me as an exercise in numerology rather than a serious scientific question. But the formation of heavy elements (beyond hydrogen) is certainly a relevant physical phenomenon that particle physics should account for (it could also be cast in terms of reaction rates and thus as a statistical phenomenon). The problem with typicality statements about fundamental constants, however, is what the relevant reference class is supposed to be. Explanations in physics end with the fundamental laws and the constants (with their specific values) are arguably part of them. Worlds with a different value of, let's say, the Higgs mass or the electron charge do not correspond to nomologically possible worlds in the sense discussed so far, namely worlds parametrized by the initial conditions for the dynamical quantities.

One could endorse a broader notion of nomic possibility, of course. But why then stop with the constants? Why not consider quantum field theories with different fields/force laws/gauge groups as representing nomic possibilities? The idea might be that physical laws should correspond somehow to mathematical *structures* – including, e.g., symmetry groups though no specific numbers – but elaborating on this would take us too far afield.

The argument that there are no meaningful *probabilities* associated with the values of the natural constants (because they are neither randomly distributed nor justifying any a priori credences) is rather uninteresting, though. Fine-tuning arguments, successful or not, should be understood in terms of typicality, not probability. Then, the relevant questions become clear.

7.3 Typicality in Mathematics

The wide scope of typicality is well illustrated by its use in "pure" mathematics. Typicality results are remarkably common in various mathematical disciplines. What seems to be lacking so far is a unified theory – or at least a more universal appreciation and consistent use – of the concept. The following discussion is a modest attempt to improve upon this situation.

One problem is the parallel use of different terminologies, which may cloud the fact that they are ultimately referring to one and the same concept. Most commonly found are statements of results which hold *almost everywhere* in some measure space, or *for almost all* members of a relevant set of points, numbers, functions, etc. More rarely, "small" exception sets are referred to as *negligible* sets. Indeed, in certain contexts, they can be literally irrelevant for the realization of some more "coarse-grained" property. For instance, changing finitely many elements of an infinite sequence does not affect its convergence properties. Measurable functions that are almost everywhere the same (i.e., $f(x) = g(x)$ for all x except for a set of measure zero) are indistinguishable by Lebesgue integration and thus identified, in the sense of equivalence classes, as elements of an L^p space. The notion of a *generic property* can also be found in the literature and is essentially a synonym for *typical property.* Formally, these notions are made precise in terms of topology, cardinality, or measure theory (see Ch. 6). In many cases, typicality results also come disguised – I believe erroneously – as probabilistic statements.

Here are some examples of typicality results in mathematics:

1. **Almost all real numbers are irrational/transcendental/uncomputable.**

   This is true in the sense of "all except for countably many" and *a forteori* also in the sense of "all except for a Lebesgue null set."

2. **Almost all real numbers are normal**, meaning that the digits $0, ..., (b-1)$ appear with equal frequency if the numbers are expanded in the integer basis b for any b. The first rigorous proof is due to (Borel, 1909).

3. **A monotone function $f : (a,b) \to \mathbb{R}$ is almost everywhere differentiable.** [Lebesgue's theorem for the differentiability of monotone functions]

4. **A bounded function $f : [a,b] \to \mathbb{R}$ is Riemann-integrable if and only if it is almost everywhere continuous.** [Riemann-Lebesgue theorem]

5. **Given m linearly independent vectors $\{v_1, ..., v_m\}$ in a vector space V of dimension $n > m$, typical vectors are linearly independent of $\{v_1, ..., v_m\}$.**

   This is true because $w \in V$ is linearly dependent, if and only if it lies in the m-dimensional subspace spanned by $\{v_1, ..., v_m\}$.

6. **A typical quadratic matrix is invertible.**

   Analogous to the previous example, it is straightforward to show that a typical array of n^2 real numbers will form linearly independent columns and thus a matrix of full rank.[2]

7. **Almost all values of a smooth map between smooth manifolds are regular values.** [Sard's theorem]

   Given a smooth map $f : M \to N$, the critical set $X \subset M$ consists of those points x at which the differential $\mathrm{d}f(x)$ has a rank $< \dim(N)$. Sard's theorem states that the image $f(X)$ – the set of critical values – has Lebesgue measure 0 in N (while X itself may be large).

8. **Khinchin's theorem**. Let $\psi : \mathbb{Z}^+ \to \mathbb{R}^+$ be a non-increasing function. A real number x is called *ψ-approximable* if there exist infinitely many rationals $\frac{p}{q}, q > 0$ such that

   $$\left| x - \frac{p}{q} \right| < \frac{\psi(q)}{q}. \tag{7.2}$$

   Khinchin (1926) proved that if the series $\sum_{q=1}^{\infty} \psi(q)$ diverges, almost all real numbers are ψ-approximable, and if the series converges, almost all real numbers are not ψ-approximable. A more general statement about such "Diophantine approximations" is the *Duffin–Schaeffer conjecture* (Duffin and Schaeffer, 1941),

[2] Moreover, topologically, the set of invertible matrices is open and dense. Open because it is the pre-image of $\mathbb{R} \setminus \{0\}$ under the continuous function det and dense because every singular matrix can be approximated by a sequence of invertible ones.

a proof of which was very recently announced (Koukoulopoulos and Maynard, 2019).

9. **Typical graphs are asymmetric.**

   Let $\Omega(n)$ be the set of (non-directed) graphs with n vertices. Then $|\Omega(n)| = 2^{\binom{n}{2}}$.

   A graph $\Gamma \in \Omega(n)$ is called *symmetric* if it has a non-trivial automorphism group, i.e., if there exists a non-identical permutation of its vertices that leaves the graph invariant.

   A famous theorem by Erdős and Rényi (1963) establishes the following (stronger) result: For $\epsilon > 0$, let $\mathcal{A}(n, \epsilon) \subset \Omega(n)$ be the set of graphs that cannot be transformed into a symmetric graph by changing at most $\frac{n(1-\epsilon)}{2}$ edges. (It is always possible to obtain a symmetric graph with at most $\frac{n-1}{2}$ changes.) Then:

   $$\lim_{n\to\infty} \frac{|\Omega(n) \setminus \mathcal{A}(n, \epsilon)|}{|\Omega(n)|} = 0.$$

10. **Every orthonormal basis is uniformly distributed over the sphere.**

    Theorem (Goldstein, Lebowitz, Tumulka, and Nino Zanghì, 2017)**.**
    Let V^n be an n-dimensional (real or complex) Hilbert space with $n \geq 4$. Let $\mathbb{S}(V^n) = \{x \in V^n : \langle x, x\rangle = 1\}$ be the unit-sphere in V^n with the uniform measure (i.e., the normalized surface area) λ. Let $G \cong \mathrm{O}(n)$ or $\mathrm{U}(n)$ the orthogonal or unitary group on V^n and μ_G the uniform (Haar) measure on G. Then, for any orthonormal basis $B = \{b_1, \dots, b_n\}$ and $\epsilon, \delta > 0$ with $n \geq \frac{1}{\delta^2 \epsilon}$

    $$\mu_G \left(\left\{ R \in G : \left| \frac{\#(B \cap R(A))}{n} - \lambda(A) \right| \leq \delta \right\} \right) \geq 1 - \epsilon, \tag{7.3}$$

    for every Borel measurable set $A \subset \mathbb{S}(V^n)$.

    This may not be so easily recognizable as a typicality result since it holds for *all* orthonormal bases rather than typical ones. Indeed, any two orthonormal bases differ by a rotation (an orthogonal or unitary transformation), so if we say that the vectors of one basis $B = \{b_1, \dots, b_n\}$ are uniformly distributed over the unit sphere, it makes sense to claim so of *every* orthonormal basis. The question is rather what it means for n discrete points on the sphere to be (approximately) uniformly distributed. And this is were Goldstein et al. invoke a typicality property: Given any (measurable) $A \subset \mathbb{S}(V^n)$ and its congruent (rotated) sets $R(A)$, $R \in G$, nearly all of them are such that the fraction of base vectors contained in $R(A)$ is approximately equal to the fraction of surface area which $R(A)$ occupies on the sphere.

It is not uncommon for mathematicians to use probabilistic language when stating some such results (less so in number theory, more so in graph theory where it leads to the study of "random graphs"). Indeed, Erdös and Rényi formulate their theorem in terms of the "probability" that a graph can be transformed into a symmetric one. Goldstein et al., while explicitly making the connection to typicality, refer to the test-sets $R(A)$ as "random rotations," and announce that any orthonormal basis in high dimensions "will pass the random test [for uniformity] with probability close to 1."

In a purely technical sense, these statements are perfectly correct, and the methods of proof may indeed draw a lot from probability theory. Conceptually, though, the reference to probability is misplaced and should be abandoned in favor of typicality.

Goldstein et al. define in terms of typical test-sets when a set of points is "uniformly distributed"; then they prove that *any orthonormal basis in high dimension is uniformly distributed over the sphere*, not that an orthonormal basis is *probably* uniformly distributed, or something like that. In particular, if one disagreed with their choice of a uniform measure on the rotation group, one would not disagree about the likelihood of finding a uniformly distributed orthonormal basis – nor even about the notion of typicality – but about the very meaning of "uniformly distributed." The concept of "random tests," briefly invoked in the abstract, may be a good heuristic, but the authors are clearly not proposing a positivist or instrumentalist notion of uniformity. Instead, they explicitly draw parallels to other "typicality theorems about spheres in high dimension," such as "most of the area of a sphere is near the equator" and "most of the volume of the unit ball is near the surface." (p. 703). Evidently, there is nothing random about the distribution of volume in a ball, nor does the truth of such propositions depend on statistical tests.

In their seminal work on asymmetric graphs, Erdös and Rényi formulate the assumption that "all possible $2^{\binom{n}{2}}$ graphs should have the same probability to be chosen" when they actually establish – in the most mundane sense of *counting* – that a certain property, viz. being asymmetric, is shared by the great majority of graphs of order n. Their theorem is, in fact, not about choosing any graph, but a mathematical fact about the sets of all finite graphs. It is, in other words, a typicality fact.

If we want to apply the theorem to a situation where a graph (or some structure isomorphic to a graph) is indeed chosen or produced – be it by a physical or maybe by a psychological process – a probabilistic language might become appropriate. But then we are, strictly speaking, outside the purely mathematical realm and need to ask the additional question, what it is about the relevant process that justifies the assumption of uniform probability. If graphs are produced by a physical process, the question becomes whether a uniform probability is typical under the pertinent physical dynamics (the relevant typicality measure for this question is then not one on the set of graphs). And how likely the woman on the street (or in your math department) is to produce a certain graph when asked for an example seems like a question for an empirical social study rather than mathematics.

Interestingly, though, there is a partially mathematical explanation for the fact that an analogous poll, asking participants to name a *real number*, may produce mostly algebraic ones despite their being atypical in $\mathbb{R}$. Since there are uncountably many transcendental numbers, almost all of them – that is, all except for a countable subset – cannot be described in any language, or expressed in a closed formula, or produced by any finite algorithm. In addition, it is very difficult to prove that a given number is not algebraic, i.e., not the root of some polynomial with rational coefficients (which is why the corresponding proof for π was such a seminal result). Hence, while almost all numbers are transcendental, we actually know almost none of them.

The bigger point here draws an interesting contrast between typicality in mathematics and in physics. We have seen that the most interesting typicality statements in physics have a modal character, referring to a reference class of possible worlds or initial conditions. This modal character is generally absent in mathematics. Physics studies the laws of nature, that is, in particular, the modal structure of the laws, but ultimately has to explain the phenomena of the actual world. Mathematics studies abstract sets and structures and one finds that many of their interesting features are typicality fact. But those typicality facts are rarely invoked to explain the properties of one particular member of the reference set since this is not the kind of problem that mathematics usually deals with.

To conclude this section, we note that typicality is a very important, yet underappreciated concept in mathematics. Conversely, the mathematical applications are particularly instructive for differentiating typicality from probability, in general. This is especially true if one shares in the Platonist intuition that there are objective typicality facts in the mathematical realm. But regardless of deeper philosophical commitments: casting internal mathematical results in terms of probabilistic models seems like a conceptual crutch at best and a category mistake at worst. I thus believe that the mathematical discourse could benefit from adopting the typicality language more consistently.

Part II

Physics

Chapter 8

From the Universe to Subsystems

> Nothing happens at random, but all things for a reason and of necessity.
>
> — Leucippus DK 80 B2

Any fundamental physical theory is a theory of the universe as a whole. Its laws describe the evolution of the *entire* configuration of matter. Thus, in classical mechanics (CM), where the forces range all over physical space, the motion of any particle at any given time depends, strictly speaking, on the positions of all the other particles and thus on the initial state of the entire universe. In quantum mechanics (QM), due to entanglement, the only fundamental quantum state is the one pertaining to the universe as a whole and represented by the universal wave function. However, this fundamental point of view is utterly impractical for everyday science, which seeks to apply these theories locally, to small parts of the physical universe. Aside from our limited computational resources, we simply do not know the exact configuration of matter and/or the exact wave function of the universe so that we could solve the equations of motion for them.

In order to derive testable propositions from a physical theory, we thus need a procedure to get from fundamental laws, describing the global evolution of the universe, to predictions about subsystems. Such a procedure was proposed by Ludwig Boltzmann, whose derivation of thermodynamic laws from microscopic particle dynamics can be viewed as a general scheme for probabilistic reasoning in the face of incomplete information. As Einstein noted, Boltzmann's insights are very much independent of the details of the underlying microscopic theory (see Einstein's *autobiographical notes* in Schilpp (1949)), and the aim of this chapter is to show how they apply to both classical and quantum mechanics – in particular, if the latter is understood in terms of Bohmian mechanics. This look ahead on quantum mechanics (to which we shall return in Chapter 12) is instructive because the Bohmian analysis of *quantum equilibrium* is, in many ways, the realization *par excellence* of Boltzmann's program. It will also allow us to address the question what, if any, is the difference between probabilities in CM and in QM.

Concerning quantum mechanics, we will focus on the theory going back to de

Broglie (1928) and Bohm (1952a) whose dominant contemporary version is known as Bohmian mechanics (BM) (Dürr et al., 2013b). The primary reason for doing so is that standard quantum mechanics runs into the infamous measurement problem illustrated by Schrödinger's cat paradox (see Ch. 12). Quantum theories that solve the measurement problem by being committed to a definite configuration of matter in physical space are known as primitive ontology theories. The wave function then has the job to describe how this configuration evolves in time, rather than to provide a complete description of the physical state. Bohmian mechanics is the most prominent example of such a theory. The primitive ontology here are particles characterized by their positions in space. The configuration of particles then evolves according to a nonlocal deterministic law of motion in which the wave function enters.

I consider Bohmian mechanics to be the most compelling non-relativistic quantum theory, not only because it provides the most obvious solution to the measurement problem but also because of the conceptual clarity and mathematical precision with which it allows us to derive the standard quantum formalism from two simple equations. Making this broad case is, however, beyond the scope of this thesis (for relevant discussions, see, e.g., Esfeld (2014a); Bricmont (2016); Norsen (2017); Dürr and Lazarovici (2018); Maudlin (2019)).

We will focus on the fact that Bohmian mechanics allows for a rigorous typicality analysis grounding the Born rule in deterministic particle dynamics and thus provides a very clear and utterly unmysterious account of probabilities in quantum mechanics. In orthodox QM, the source of randomness is the measurement process and the collapse of the wave function superseding the deterministic Schrödinger evolution. But since measurements are treated as primitive, and the collapse postulate is obscure (if not plainly inconsistent), the status of this randomness remains obscure, as well. Bohmian mechanics, in contrast, provides the clearest counterexample to the widespread belief that "quantum probabilities" must be fundamentally different from the probabilities used in classical mechanics.

Nonetheless, there appear to be striking differences between CM and QM that we have to account for. In particular, also in Bohmian mechanics, one cannot do better than making statistical predictions according to Born's rule. In CM, by contrast, we are used to dealing with situations in which we can obtain a reliable deterministic description of certain subsystems. We will explain why this is so, and thus address the bigger question of why quantum mechanics *appears* more stochastic than classical mechanics.

8.1 Probabilities in Classical Mechanics

In classical mechanics, the physical state of an N-particle system is completely determined by specifying the positions and momenta of all the particles. Denoting by q_i and p_i the position, respectively the momentum, of the i'th particle, we call

$X(t) = (q_1(t), ..., q_N(t); p_1(t), ..., p_N(t))$ the *microstate* of the system at time t. The space of all possible microstates, here $\Gamma := \mathbb{R}^{3N} \times \mathbb{R}^{3N}$, is the *phase space*. The microstate evolves according to the microscopic laws of motion, which, in the Hamiltonian formulation, take the form

$$\begin{cases} \dot{q}_i = \frac{\partial H}{\partial p_i} \\ \dot{p}_i = -\frac{\partial H}{\partial q_i} \end{cases}, \tag{8.1}$$

with

$$H(q,p) = \sum_{i=1}^{N} \frac{p_i^2}{2m_i} + V(q_1, \ldots, q_n). \tag{8.2}$$

More compactly, this can be written as

$$(\dot{q_i}, \dot{p_i}) = v^H(q,p), \tag{8.3}$$

where v^H denotes the vector field on Γ generated by the Hamiltonian H. These equations give rise to a Hamiltonian flow $\Phi_{t,0}$ such that $X(t) = \Phi_{t,0}(X)$ for any initial microstate X at $t = 0$. In equation (8.2), m_i denotes the mass of the i'th particle and V the interaction potential, which can be split into

$$V(q_1, \ldots, q_n) = \sum_{i<j} V_{int}(q_i - q_j) + V_{ext}(q_1, ..., q_N, t). \tag{8.4}$$

V_{int} then corresponds to a pair-interaction among the particles (e.g., gravitation) and V_{ext} is an external potential summarizing the influences of the environment. Of course, if the N particle system is the entire universe, then $V_{ext} = 0$ since there is nothing outside the universe.

Notably, whenever we argue that a subsystem can be treated as isolated, i.e., that V_{ext} is negligible, for instance, because of the large distance/small mass of other bodies in Newtonian gravity, we are assuming the universal validity of the laws. The relevant argument doesn't go bottom-up – to larger and larger "models" of the theory – but top-down from the universe to subsystems.

In any case, a Hamiltonian system has several nice properties. If V_{ext} is zero, or at least time-independent, it conserves the total energy, meaning that $H = const.$ along any solution of (8.1). Furthermore, by the Liouville theorem, the Hamiltonian flow conserves phase space volume. This is to say that the uniform Lebesgue measure λ is a *stationary* measure on Γ, in the sense that for all $t \geq 0$ and any Borel set $A \subseteq \Gamma$,

$$\lambda(\Phi_{t,0}A) = \lambda(A). \tag{8.5}$$

For fixed $E \in \mathbb{R}$, one will usually consider the reduced phase space $\Gamma_E := \{X \in \Gamma : H(X) = E\}$ to which a system with total energy E is confined by virtue of energy conservation. λ then induces a stationary measure λ_E on the hypersurface Γ_E, called the *microcanonical measure*. By convention, we normalize this measure to $\lambda_E(\Gamma_E) = 1$.

Remark. Stationarity and Continuity Equation.

In general, a measure μ evolves under a flow $\Phi_{t,0}$ as

$$\mu_t(A) = \mu\left(\Phi_{t,0}^{-1}(A)\right), \tag{8.6}$$

i.e., $\mu_t(A)$ corresponds to the measure that $\mu = \mu_0$ assigns that the points that have evolved into the region A at time t. This means that for any test-function f:

$$\int f(x)\, \mathrm{d}\mu_t(x) = \int f(\Phi_{t,0}(x))\, \mathrm{d}\mu(x). \tag{8.7}$$

If $\mu = \rho\, \mathrm{d}x$ with a density ρ, then μ_t has a density $\rho_t = \rho(t, x)$ such that, by (8.7),

$$\int f(x)\rho(t,x)\, \mathrm{d}x = \int f\left(\Phi_{t,0}(x)\right)\rho(x)\, \mathrm{d}x. \tag{8.8}$$

The stationary condition requires that these integrals are actually constant in t, thus

$$\begin{aligned} 0 &= \frac{\mathrm{d}}{\mathrm{d}t}\int f\left(\Phi_{t,0}(x)\right)\rho(x)\mathrm{d}x = \int \rho(x)v\left(\Phi_t(x)\right)\cdot\nabla f\left(\Phi_{t,0}(x)\right)\mathrm{d}x \\ &= \int \rho(x,t)\, v(x)\cdot\nabla f(x)\, \mathrm{d}x = -\int f(x)\, \mathrm{div}(v(x)\rho(x,t))\, \mathrm{d}x, \end{aligned}$$

where the last equality follows by partial integration (assuming that f falls off quickly to zero towards infinity). Since this must be valid for any test-function f, we read off the *continuity equation*

$$\partial_t\rho(x,t) + \mathrm{div}\left(v(x)\rho(x,t)\right) = 0. \tag{8.9}$$

Informally speaking, the equation says that weight doesn't get lost but is only transported along the flow-lines.

For a Hamiltonian vector field, it is easy to check (from eqs. (8.1) and (8.3)) that $\mathrm{div}\, v^H = 0$, so that $\rho = const.$ is a solution of (8.9). If one wants to be mathematically fancy about it, one can relate this to the *symplectic structure* of phase space, which ensures that the canonical volume form $\mathrm{d}x = \mathrm{d}q \wedge \mathrm{d}p$ does not change under the Hamiltonian flow. In general, however, a stationary *measure* is not equivalent to a stationary *density* since the volume form may itself evolve with the dynamics.

Randomness and typicality

Given that CM is deterministic, where do probabilities enter the picture? From a practical point of view, there are at least three reasons to depart from the deterministic description: (i) we do not have access to the exact values of all positions and momenta in a given physical system. We can neither manipulate them with arbitrary precision in experimental situations, nor measure the exact (initial) microstate X in order to determine the system's trajectory. (ii) Physical systems can be extremely sensitive to perturbations of their initial conditions. This means that even a small error about the initial data can translate into a huge error about the evolution of the system. (iii) The complexity of calculation increases rapidly as N becomes very large.

Against this background, it seems reasonable and necessary to make two concessions. First, it usually suffices to provide a *coarse-grained* description of the system. That is, rather than asking for the exact microstate, we are interested in the value of certain macroscopic "observables" $F : \Gamma \to \mathbb{R}$. These observables are coarse-graining in the sense that a great number of microstates X will, in general, correspond to (approximately) the same value of F. Mathematically, if Γ is endowed with a probability measure, such a function is called a *random variable*, but as emphasized before, this nomenclature is quite deceiving. The *macrostate* of a system (defined in terms of such observables) is always determined by its microstate which, in turn, follows a deterministic law of motion.

Second, although we cannot determine the exact evolution of the system – if only for the fact that we do not know the exact initial conditions – we can ask what happens in *most* possible instances, that is, for *typical* initial conditions.

In some cases, typical trajectories coarse grain to one and the same macroscopic history, so that predictions appear deterministic (e.g., when we set out to determine the trajectory of a stone thrown on earth). In many cases, though, typical initial conditions agree only on certain statistical patterns in the distribution of coarse-grained observables (e.g., when we ask for the relative frequency of *heads* or *tails* in a long series of coin tosses). In these latter cases, probabilities come into play.

In any case, if we can establish that a certain fact or feature occurs for the vast majority of possible initial conditions – that is, in the last resort, for the overwhelming majority of possible universes described by a particular theory –, we can justifiably call it a prediction of that theory. In order to make such an argument precise, we need a measure on phase space – a typicality measure – telling us what an "overwhelming majority" of initial conditions is.

In CM, the natural typicality measure is the Lebesgue or Liouville measure on phase space, respectively the induced microcanonical measure on the energy shell. We have already discussed how this particular choice is justified and what characterizes a good typicality measure, in general. Here, we shall highlight again *stationarity* as a crucial desideratum, since it is essential to a sensible notation of typicality that it does not change with time. Stationarity of the measure, i.e., equation (8.5), assures that typical sets remain typical, and atypical sets remain atypical under the time evolution. In Hamiltonian mechanics, the uniform Liouville measure is thus distinguished as the simplest stationary measure on phase space, and when we come to Bohmian QM, we shall see that it is indeed the stationary measure – not the uniform measure – that yields the appropriate notion of typicality.

Through the condition of stationarity, the choice of typicality measure is constrained by the physical dynamics. However, as noted in Ch. 5.5, stationarity alone may not distinguish the measure uniquely. In fact, for classical mechanics, any density $f(H)$ that is function of the Hamiltonian H defines a stationary measure on phase space. I have already argued that the additional appeal to more informal criteria such as simplicity and naturalness is quite appropriate. Nonetheless, the situation is more satisfying in Bohmian mechanics, where the typicality measure can be shown to be unique in a rather strong sense.

Ideal Gas: The Maxwell distribution

To demonstrate how a typicality argument works, let us consider the stock example of an ideal gas in a box (with perfectly reflecting walls) that will serve as our toy-model for the universe. The number of particles in such a macroscopic system is of the order of Avogadro's constant, which is $N \sim 10^{24}$. Clearly, determining the actual configuration and / or predicting the trajectories for so many particles is a hopeless task, even if the particles are non-interacting as in our example.

And yet, it is possible to make reliable predictions about the system. For instance, we can ask the following: what is the share of particles whose velocity in x-direction is approximately $v_0 \in \mathbb{R}$? We can formalize this in terms of the random variable

$$F(X) := \frac{1}{N}\sum_{i=1}^{N} \chi_{\{v_{i,x}\in[v_0-\delta,v_0+\delta]\}}(X). \tag{8.10}$$

Here, $\delta > 0$ is a small positive number (giving precise meaning to "approximately v_0") and χ is the indicator function, i.e., $\chi_{\{v_{i,x}\in[v_0-\delta,v_0+\delta]\}}$ equals one if $v_{i,x} = \frac{1}{m}p_{i,x}$ lies in the interval $[v_0 - \delta, v_0 + \delta]$ and zero if it does not.

Fixing the mean energy per particle to $\frac{E}{N} = \frac{3}{2}k_{\mathrm{B}}T$ (k_{B} is the Boltzmann constant and T can later be identified as the temperature of the system), it is a mathematical fact that for any $1 \leq i \neq j \leq N$:

$$\begin{aligned}\lim_{N\to\infty,\frac{E}{N}=\frac{3}{2}k_{\mathrm{B}}T} & \lambda_E\Big(\Big\{X \in \Gamma_E : v_{i,x} \in [a,b], v_{j,x} \in [c,d]\Big\}\Big) \\ &= \int_c^d \int_a^b \frac{\exp\left(-\frac{1}{k_{\mathrm{B}}T}\frac{m(v_i^2+v_j^2)}{2}\right)}{\left(\frac{2\pi k_{\mathrm{B}}T}{m}\right)^3}\mathrm{d}v_i\mathrm{d}v_j\,.\end{aligned} \tag{8.11}$$

From this result – notably establishing (pairwise) statistical independence in the thermodynamic limit – one can conclude that for any $\epsilon > 0$:

$$\lambda_E\Big(\Big\{X : \Big|\frac{1}{N}\sum_{i=1}^{N}\chi_{\{v_{i,x}\in[a,b]\}}(X) - \int_a^b \rho_{MB}(v)\mathrm{d}v\Big| > \epsilon\Big\}\Big) \to 0,\ N \to \infty, \tag{8.12}$$

with

$$\rho_{MB}(v) = \Big(\frac{2\pi k_{\mathrm{B}}T}{m}\Big)^{-\frac{3}{2}} \exp\left(-\frac{mv^2}{2k_{\mathrm{B}}T}\right). \tag{8.13}$$

The rigorous derivation requires little more than standard calculus and measure theory. The deeper philosophical question, is what the mathematical result means.

The function $\rho_{MB}(v)$ is called the *Maxwell* or *Maxwell-Boltzmann distribution*. It is a probability density, describing a distribution of particle velocities. Note that there is nothing intrinsically random about the velocities of particles in a gas. The velocity (as well as the position) of every single particle is comprised in the microstate X whose evolution follows a deterministic equation of motion. There are possible X for which the actual distribution of velocities in the gas differs significantly from that described by the Maxwell distribution. For instance, there are microstates X for which all particles move with one and the same velocity, or microstates X for which a few very fast particles account for almost the entire kinetic energy while all the others are nearly at rest. But these states are very special ones. The crucial and remarkable fact expressed by equation (8.12) is that, for large N, the *overwhelming majority* of possible microstates is such that the distribution of velocities in the gas is (approximately) Maxwellian. What constitutes an “overwhelming majority of microstates” is made precise in terms of the stationary measure λ_E. The Maxwell distribution is thus derived from the microscopic theory as a statistical regularity manifested for *typical* microconfigurations.

Ludwig Boltzmann expressed this reasoning as follows:

> The ensuing, most likely state [...] which we call that of the Maxwellian velocity distribution, since it was Maxwell who first found the mathematical expression in a special case, is not an outstanding singular state, opposite to which there are infinitely many more non-Maxwellian velocity-distributions, but it is, to the contrary, distinguished by the fact that by far the largest number of possible states have the characteristic properties of the Maxwellian distribution, and that compared to this number the amount of possible velocity-distributions that deviate significantly from Maxwell’s is vanishingly small. (Boltzmann, 1896a, p. 252, translation by the author)

It is crucial to appreciate that while two – actually even three – measures appear in the mathematical expression (8.12), their status is very different (see Goldstein (2012)). We have:

- The actual (empirical) distribution $\rho_{emp}[X]\,\mathrm{d}v = \frac{1}{N}\sum\limits_{i=1}^{N}\chi_{\{v_{i,x}\in \mathrm{d}v\}}(X)$, yielding, for the microstate X, the fraction of particles with x-velocity in the interval $\mathrm{d}v$.
- The theoretical (Maxwellian) distribution $\rho_{MB}\,\mathrm{d}v \propto \exp\left(-\frac{1}{k_{\mathrm{B}}T}\frac{mv^2}{2}\right)\mathrm{d}v$.
- And the typicality (microcanonical) measure λ_E.

Equation (8.12) thus tells us that $\rho_{emp} \approx \rho_{MB}$ for *typical* microstates $X \in \Gamma_E$. Notably, the Maxwellian ρ_{MB} and the empirical distribution ρ_{emp} refer to the ensemble of particles within the box, whereas the microcanonical measure does not refer to an ensemble of boxes but is used to define typicality.

I also want to emphasize again the very limited degree to which *knowledge*, *information*, *credence* or other subjective notions play a role in the analysis. While it is in some sense correct to say that randomness in a deterministic theory is only due to our ignorance about initial conditions, it is an objective fact that for the great majority of microstates the distribution of velocities in an ideal gas is (approximately) Maxwellian. It is this objective fact, rather than some quantification of our knowledge or beliefs, that we take to be explanatory.

The coin toss again

Analogous reasoning can be applied to more mundane examples like the repeated tossing of a coin. It is a statistical regularity found in our universe that the relative frequency of heads or tails in a long series of fair coin tosses is approximately 1/2. Since coin tossing is guided by the same laws as all other physical processes in the world, this statistical regularity has to be explained on the basis of the fundamental microscopic theory (here: classical mechanics). It is not a new kind of law that holds over and above the microscopic ones.

We have already seen what such an account would look like. We denote by $\chi_i(X) \in \{0, 1\}$ the outcome of the i'th coin toss in a long series of N tosses. Since classical mechanics is deterministic, the outcome of every single trial is determined, through the fundamental laws of motion, by initial conditions X. Obviously, the functions χ_i are very coarse-graining. We do not care about the exact configuration of atoms making up the coin; we do not even care about the exact position or orientation of the coin. We only ask which side is facing up as the coin lands on the floor. This defines our macroscopic observables.

It is not wrong to think of X, in the first instance, as ranging over possible initial configurations of the before-mentioned coin-tossing machine: At time $t = 0$, a large number N of coins is filled into the machine, which is then sealed and shielded from outside influences. From there on, everything takes its deterministic course: the outcome of each coin toss is completely determined by the initial state of the system. However, the account would remain incomplete, not just because a perfectly isolated machine is an unrealistic idealization. The initial configuration of the coin-tossing machine is itself the result of physical processes (the process of setting up the machine, for instance) that are determined by suitably specified initial conditions. And these initial conditions are the result of other deterministic processes in even larger systems, and so on and so forth. If we think this trough till the end, we must eventually speak about the universe as a whole – the only truly closed system – and think of χ_i as functions on the phase space of the universe (respectively, the Past Hypothesis macro-region).

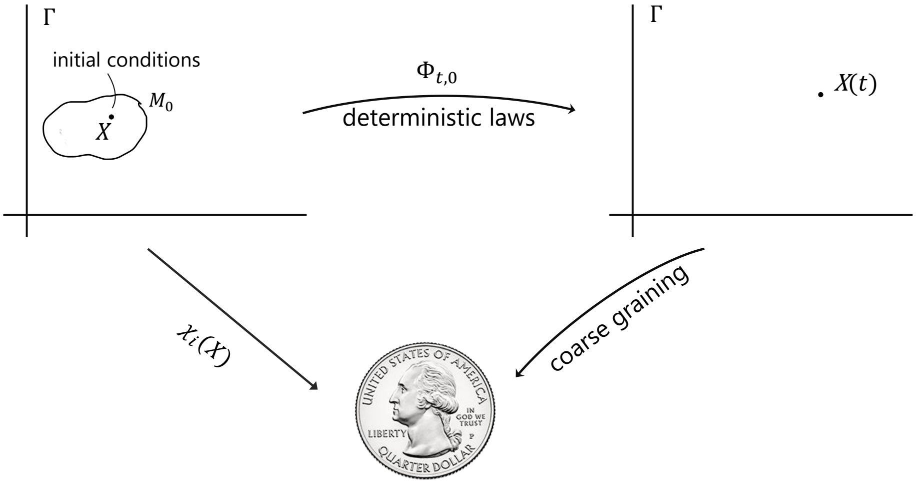

Figure 8.1: Sketch: a macroscopic event supervening on the microscopic evolution. $\Phi_{t,0}$ is the flow arising as the general solution of the microdynamics.

We know, of course, that classical mechanics are not adequate everywhere and on all scales throughout our universe. Nevertheless, if we want to argue that the Newtonian theory is sufficient to explain coin toss statistics – and it is hard to see why relativistic or quantum effects should be relevant – we must ultimately conceive of these statistics as a regularity in a Newtonian universe.

Now, there are possible initial configurations that would give rise to a Newtonian universe in which the relative frequency of heads is very different from 1/2. Very plausibly – that is, if the usual assumption of statistical independence is somewhat justified – there are possible initial conditions for which almost *all* coins ever tossed land on heads, or for which 2 out of 3 tosses result in tails, and so on and so forth. But such initial conditions are very special ones. In contrast, typical initial conditions (compatible with there being coins and coin tossers in the first place) are such that the relative frequency of heads or tails in a long series of trials is approximately 1/2. More formally, the claim is that for any $\epsilon > 0$,

$$\lambda\Big(\Big|\frac{1}{N}\sum_{i=1}^{N}\chi_i(X) - \frac{1}{2}\Big| > \epsilon \;\Big|\; M_0\Big) = \delta(\epsilon, N), \tag{8.14}$$

where M_0 is the initial macrostate of the coin-tossing machine or, better, the universe (constrained by the relevant facts about the experiment) and $\delta(\epsilon, N)$ becomes arbitrarily small with increasing N. This is to say that if N is sufficiently large, the set of initial conditions for which the relative frequency of heads deviates significantly from 1/2 is extremely small. Such initial conditions are not *impossible* but *atypical.*

By now, the reader will have certainly made the connection to our previously-discussed model in terms of the Rademacher functions and also recognized (8.14) as a law of large numbers statement.

Deterministic subsystems: the stone throw

As mentioned before, there are many situations in classical mechanics that seem very different from the coin toss or the velocity distribution of molecules in a gas. For instance, when we predict the trajectory of a stone thrown on earth, we can, in general, use a simple deterministic equation without being embarrassed by our ignorance regarding the exact initial microstate of the stone or its environment. There are two conditions satisfied here that allow us to do so:

1. The external forces, that is, the influence of the rest of the universe neglected in our calculations, is very small compared to the attraction between stone and earth. This is because other gravitating bodies are either very far away or have very small masses compared to our planet. Formally, this is to say that

   $$V_{ext} \approx 0, \tag{8.15}$$

   which allows us to treat the system (stone, earth) for most practical purposes as an isolated Newtonian system.

   We emphasize again that such arguments presuppose the universal validity of the fundamental law. When we argue that the gravitational attraction of Uranus has a negligible influence on the trajectory of the stone, we are obviously accepting that the law of gravitation applies to planet Uranus.

2. The evolution of the relevant macroscopic variable – here, the center of mass of the stone – is reasonably robust against variations of the microscopic initial conditions. In other words, small changes in the initial micro-conditions have (typically) a small effect on the trajectory of the stone. This is why our ignorance about the exact microscopic configuration of the stone (or of the planet earth, or the person/apparatus throwing the stone) does not prevent us from making reliable predictions about the motion of its center of mass.

Nonetheless, even in this case, our prediction will be strictly speaking a typicality result. Atypical events in the environment or the many-particle system constituting the stone can lead to very different outcomes. To be precise, we would actually have to cast our mechanical prediction for the stone's trajectory in a form that looks very similar to the probabilistic statements (8.12) or (8.14). For instance, denoting by $x(t)$ the computed trajectory (depending on the initial position and momentum of the stone) and by $\tilde{x}(t)$ the actual trajectory of the stone (depending on the initial condition X of the universe), we could write:

$$\lambda\left(\left\{X : \sup_{t\in[0,T]} |\tilde{x}(t) - x(t)| > \epsilon\right\}\middle| M_0\right) \approx 0, \tag{8.16}$$

where the macrostate M_0 includes our approximate (coarse-grained) knowledge of the initial conditions, as well as our evidence justifying our description of the system (stone, earth) as an isolated Newtonian subsystem.

That said, the stone throw example still points to a striking difference between classical and quantum mechanics. In CM, we routinely deal with situations in which correlations between a subsystem and its environment are irrelevant and deterministic predictions for the subsystem prove to be successful. In QM, by contrast, we are generally dealing with situations that are much more similar to the coin toss or the molecules in a gas, where predictions of statistical patterns are the best we can hope for. Our aim now is to explain why this is so. To this end, we first have to discuss probabilities in the quantum case.

8.2 Probabilities in Bohmian Mechanics

> Assuming the success of efforts to accomplish a complete physical description, the statistical quantum theory would, within the framework of future physics, take an approximately analogous position to the statistical mechanics within the framework of classical mechanics. I am rather firmly convinced that the development of theoretical physics will be of this type; but the path will be lengthy and difficult.
>
> — Albert Einstein in (Schilpp, 1949, p. 672)

In QM, we encounter a new dynamical feature that is totally absent from CM: the specification of initial positions and momenta is replaced with the specification of an initial wave function. The wave function is defined on configuration space and, in general, non-separable. That is, due to *entanglement*, wave functions cannot be attributed to the particles individually as initial parameters were attributed to them individually in CM. On the fundamental level, there exists only one wave function, the *universal wave function*, pertaining to the whole particle configuration of the universe taken together.

As announced above, we shall consider the precise quantum theory known as Bohmian mechanics (BM). BM is characterized by the following three postulates:

1. A Bohmian system with N particles is completely described by a pair (Q, Ψ), where $Q = (Q_1, \ldots, Q_N) \in \mathbb{R}^{3N}$ represents the spatial configuration of the particles and Ψ is a complex square-integrable function on the configuration space $\mathbb{R}^{3N}$ called the universal wave function.

2. The evolution of the wave function Ψ is described by the Schrödinger equation

$$i\hbar\partial_t\Psi_t = H\Psi_t, \tag{8.17}$$

   where H is the Hamiltonian of the system.

3. The evolution of the particle configuration follows a first-order differential equation in which the wave function Ψ_t enters to determine a velocity field v^{Ψ_t} for

> the particles. More precisely, the particle configurations evolves according to the *guiding equation*
>
> $$\dot{Q} = v^{\Psi_t}(Q) := \frac{\hbar}{m}\mathrm{Im}\frac{\nabla\Psi_t(Q)}{\Psi_t(Q)}, \tag{8.18}$$
>
> where we assume, for simplicity, particles of equal mass m, ∇ is the gradient on the $3N$-dimensional configuration space, and Im denotes the imaginary part. Note that, due to the entanglement of the wave function, the law of motion is manifestly nonlocal: except for special situations, the velocity of a particle will depend on the position of all the other particles at the same time.

Given an initial wave function Ψ_0 and the initial particle configuration $Q_0 \in \mathbb{R}^{3N}$, the evolution of the system is completely and uniquely determined for all times. This determinism is contrary to the popular belief that quantum mechanics is intrinsically and irreducibly random. However, since we do not know (in fact, as we will see, *cannot* know) the exact particle configuration, we have to resort once again to a statistical analysis in order to extract empirical predictions. To this end, we can pursue the same strategy as we did before in CM.

In the following, we will largely rely on the development of this strategy by Dürr et al. (1992) (repreinted as Ch. 2 in Dürr et al. (2013b); for a textbook discussion, see Dürr and Teufel (2009)). For the statistical analysis of BM, we need a) a sensible typicality measure defined on configuration space and b) a procedure to get from the fundamental, universal description in terms of the universal wave function to a well-defined description of Bohmian subsystems. Given the universal wave function, the appropriate notion of typicality for particle configurations is defined in terms of the measure with density $|\Psi|^2$. The crucial feature of this measure is that it is *equivariant*, assuring that typical sets remain typical and atypical sets remain atypical under the Bohmian time-evolution. More precisely, if $\Phi^{\Psi}_{t,0}$ is the flow on configuration space induced by the guiding equation (8.18), then

$$\mathbb{P}^{\Psi}(A) := \int\limits_{A} |\Psi_0|^2 \, \mathrm{d}^{3N}q = \int\limits_{\Phi^{\Psi}_{t,0}(A)} |\Psi_t|^2 \, \mathrm{d}^{3N}q \tag{8.19}$$

holds for any measurable set $A \subseteq \mathbb{R}^{3N}$. Equivariance is thus the natural generalization of stationarity for non-autonomous (time-dependent) dynamics. The $|\Psi|^2$-measure can be proven to be the unique equivariant measure for the Bohmian particle dynamics that depends only locally on Ψ or its derivatives (Goldstein and Struyve, 2007). In this sense, it is even more strongly suggested as the correct typicality measure for BM than the Liouville measure is in CM.

Let us now have a closer look at how BM treats subsystems of the universe. Suppose that the subsystem consists of $n < N$ particles. We then split the configuration space into $\mathbb{R}^{3N} = \mathbb{R}^{3n} \times \mathbb{R}^{3(N-n)}$, so that, writing $q = (x, y)$, the x-coordinates describe the degrees of freedom of the subsystem and the y-coordinates describe the possible configurations of its environment, i.e., the rest of the universe. Analogously, we split the *actual* particle configuration into $Q = (Q^{sys}, Q^{env}) = (X, Y)$, with $Q^{sys} = X$, the configuration of the subsystem under investigation and $Q^{env} = Y$ the configuration of its environment.

Now, in passing from the fundamental universal theory to a description of the subsystem, we can simply take the universal wave function $\Psi_t(q) = \Psi_t(x, y)$ and plug into the y-argument the actual configuration $Y(t)$ of the rest of the universe. The resulting

$$\psi_t^Y(x) := \Psi_t(x, Y(t)) \tag{8.20}$$

is now a function of the x coordinates only, called the *conditional wave function.* In terms of this conditional wave function, the equation of motion for the subsystem takes the form

$$\dot{X}(t) = \frac{\hbar}{m}\mathrm{Im}\frac{\nabla_x \psi_t^Y(x)}{\psi_t^Y(x)}\bigg|_{x=X(t)} \tag{8.21}$$

to be compared with (8.18).

Since the conditional wave function depends explicitly on $Y(t)$, its time-evolution may be extremely complicated and not follow any Schrödinger-like equation. However, in many relevant situations, the subsystem will dynamically decouple from its environment. We say that the subsystem has an *effective wave function* φ if the universal wave function takes the form

$$\Psi(x, y) = \varphi(x)\chi(y) + \Psi^{\perp}(x, y), \tag{8.22}$$

where χ and $\Psi^{\perp}$ have macroscopically disjoint y-supports and $Y \in \operatorname{supp}\chi$, so that, in particular, $\Psi^{\perp}(x, Y) = 0$ for almost all x. (This is, notably, much weaker than assuming that Ψ has a product structure, which is almost never justified.) This means that we can forget about the empty wave packet $\Psi^{\perp}(x, y)$ and describe the subsystem in terms of its own independent wave function φ (normalized to $\int |\varphi(x)|^2 \mathrm{d}^{3n}x = 1$). If we can furthermore assume that the interaction between subsystem and environment is negligible for some time, that is,

$$V_{ext}(x, y)\varphi(x)\chi(y) \approx 0, \tag{8.23}$$

the effective wave function will satisfy its own autonomous Schrödinger evolution. Note that from the point of view of the subsystem, this part of the interaction potential, coupling x and y degrees of freedom, is an external potential; condition (8.23) is thus the same as (8.15) above.

Effective wave functions are the Bohmian counterparts of the usual wave functions in textbook QM, and those wave functions to which the Born rule generally refers.

For our statistical analysis, we start by considering the conditional measure

$$\mathbb{P}^{\Psi}(X \in \mathrm{d}^{3n}x\} \mid Y) = \frac{|\Psi((x,Y))|^2 \mathrm{d}^{3n}x}{\int |\Psi((x,Y))|^2 \mathrm{d}^{3n}x} = |\psi^Y(x)|^2 \mathrm{d}^{3n}x, \tag{8.24}$$

where the conditional wave function ψ^Y is now normalized (and we keep in mind that in the special situations described by (8.22), it becomes an *effective* wave function). This formula already holds a deep insight to which we shall soon return. For practical purposes, though, conditioning on the configuration Y is much too specific, since we have only very limited knowledge of Y. However, many different Ys will yield one and the same conditional/effective wave functions for the subsystem. Collecting all those Ys, and using the fact that by yielding the same conditional wave function they also yield the same conditional measure (8.24), a simple identity for conditional probabilities yields

$$\mathbb{P}^{\Psi}(X \in \mathrm{d}^{3n}x\} \mid \psi^Y = \varphi) = |\varphi|^2 \mathrm{d}^{3n}x. \tag{8.25}$$

From this formula, we can now derive law of large numbers estimates of the following kind: at a given time t, consider an ensemble of M identically prepared subsystems with effective wave function φ.[1] We denote by X_i the actual configuration of the i'th subsystem. Let $A \subseteq \mathbb{R}^{3n}$ consider the corresponding indicator function $\chi_{\{X_i \in A\}}$, which is 1, if the configuration X_i is in A and 0 otherwise. Then, we have for any $\epsilon > 0$:

$$\mathbb{P}^{\Psi_t} = \Big(\Big\{Q : \Big|\frac{1}{N}\sum_{i=1}^{N} \mathbb{1}\{X_i \in A\}(Q) - \int_A |\varphi(x)|^2\Big| < \epsilon\Big\}\Big) \to 0, N \to \infty. \tag{8.26}$$

This is to say that for *nearly all* possible configurations of the universe, the particles in an ensemble of subsystems with effective wave function φ will be distributed according to $|\varphi|^2$. In other words, Born's rule holds in typical Bohmian universes, that is, in *quantum equilibrium.*

Once again, we emphasize that the $|\Psi|^2$-measure given in terms of the *universal* wave function is only used to define typicality. It is *not* supposed to describe an actual distribution of configurations (an "ensemble of universes," whatever that would mean), because the universe exists only once. By contrast, the $|\varphi|^2$-measure on the right hand side, defined in terms of the effective wave function, does refer to actual particle distributions in a typical ensemble of identically prepared subsystems. Born's rule is thus predicted and explained by BM as a statistical regularity of typical Bohmian universes.

Comparing equation (8.25) to (8.12) (and recalling the reasoning that lead to the respective equations), we recognize the analogy between the derivation of the Maxwell distribution in CM and the Born distribution in Bohmian mechanics. In essence, it is Boltzmann's statistical mechanics applied to two different theories. The status

[1] The analysis for "time-like ensembles," i.e., consecutive measurements on the same system, is mathematically more involved and carried out in Dürr et al. (1992).

of probabilities and the role of typicality is the same in both cases, although the dynamical laws are strikingly different. On the one hand, this illustrates the deepness and universality of Boltzmann's insights. On the other hand, it shows that there is no need to look for a fundamentally new kind of randomness in the quantum realm. If the microscopic laws and the ontology of the theory are clear, probabilities in QM are no more mysterious than they are in CM.

Why Quantum mechanics appears more random

We have highlighted the similarities between the statistical analysis of classical mechanics and Bohmian mechanics, showing that probabilities have the same status in both theories. But what then is the difference between the classical and the quantum regime? Why does the latter appear so much more unpredictable and random?

Part of the answer is trivial. Quantum mechanics is primarily used to describe microscopic systems, while Newtonian mechanics is successfully applied on macroscopic scales. Macroscopic predictions are bound to be more robust against our ignorance about the micro-conditions. Of course, we think of QM as more fundamental theory from which classical mechanics emerge in an appropriate "classical limit." In Bohmian mechanics, this refers to situations in which the Bohmian trajectories look approximately Newtonian on macroscopic scales (see, e.g., Dürr and Teufel (2009, Ch. 9)). This means, however, that the successful "deterministic" predictions of classical mechanics are also – and more fundamentally – predictions of Bohmian mechanics.

Furthermore, we have seen that the predictability of a system depends on our abilitiy to describe it, at least effectively, as independent of external influences. The nonlocality of quantum mechanics makes this a particularly subtle issue. Newtonian gravity is also nonlocal, but only in a milder sense. Forces fall off quickly with increasing distance (and gravity is very weak, to begin with) so that parts of the universe can often be described as autonomous Newtonian systems for all practical purposes.

Quantum mechanics, by contrast, has a distinctly holistic character. In BM, this is manifested in nonlocal dynamics, in which the entire configuration of particles is guided by a common wave function. Quantum entanglement (or what Maudlin (2011) calls the "quantum connection") is universal and does not fall off with distance. This makes it much more difficult to consider subsystems as isolated while ignoring the influence of the rest of the universe. Fortunately, many relevant situations allow for an autonomous Bohmian description of a subsystem in terms of an effective wave function. Einstein's worry (see Einstein (1948)) that nonlocality would make the investigation of nature by local experiments impossible did not manifest. Nonetheless, we must be careful since the effective wave function depends implicitly on the environment configuration (e.g., on the procedure used to prepare the state in an experiment) via equation (8.22).

Absolute Uncertainty

More precisely (and more profoundly), the information that we can possess about the configuration of a Bohmian subsystem is restricted by the theorem of *Absolute Uncertainty* (Dürr et al., 1992) that has no analog in classical physics. "Information" here is understood very prosaically as a correlation between the configuration of the subsystem and the configuration of some other system – a brain, a measurement device, a notebook – that constitutes a *record.* Absolute uncertainty is then a direct consequence of the conditional probability formula (8.24): all external records about the subsystem are included in the particle configuration Y of the rest of the universe and thus already taken into account (i.e., conditionalized on) in equation (8.24) that yields Born's rule for the distribution of particle positions.

The theorem of absolute uncertainty thus states that if the effective (or conditional) wave function of a subsystem is φ, an external observer cannot have more information about the particle configuration of that system than provided by the $|\varphi|^2$-distribution.

Conversely, this means that if we perform additional measurements to determine the particle positions with greater accuracy, the system's effective wave function must be affected and become more and more peaked. Hence, the gradients in the velocity formula (8.18) induce higher and higher possible velocities, depending on the exact particle configuration Q, and also an ever greater variation ($\sim \nabla^2\varphi$) in the possible velocities. Less uncertainty about the (initial) particle positions thus implies more uncertainty about the (asymptotic) velocities – this is the source of Heisenberg's uncertainty principle. Consequently, even tiny variations in the initial data may lead to large deviations of the corresponding Bohmian trajectories. (Our rapidly increasing uncertainty about the particle positions is then mirrored by the quick spreading of the wave function under the Schrödinger time evolution.)

Absolute uncertainty is a feature of *quantum equilibrium.* And this is maybe the deepest explanation for our epistemic limitations in the microscopic realm: Our universe is in quantum equilibrium but macroscopically in (thermodynamic) non-equilibrium. And non-equilibrium is what allows for more informative correlations. Just as a gas in equilibrium is always Maxwell-distributed (with different temperatures and/or effective Hamiltonians), Bohmian subsystems are always Born-distributed (with different effective wave functions).

In conclusion, the "randomness" of quantum mechanics is a result of quantum equilibrium and manifestly nonlocal dynamics, which are such that a system becomes immediately more chaotic as we try to determine the micro-conditions with greater accuracy. This forces us to resort much more routinely to probabilistic reasoning than is the case in classical physics. For a quantum system, Born's rule provides – provably – as good a description as we can get in a universe in quantum equilibrium.

Chapter 9

Boltzmann's Statistical Mechanics

In this chapter, we will discuss Ludwig Boltzmann's statistical mechanics and revisit the groundbreaking insights of the Austrian physicist who, more than a century ago, showed how to derive and explain macroscopic regularities on the basis of the underlying laws governing the motion of the microscopic constituents of matter. We will focus, in particular, on his account of thermodynamic irreversibility and the second law of thermodynamics in which the concept of typicality plays a central role.

In the philosophical literature, the account is sometimes referred to as "Neo-Boltzmannian" but this name strikes me as overly flattering to both Boltzmann's critics and his contemporary defenders. This is not to diminish the important contributions of, among others, Lebowitz (1993a,b), Bricmont (1995), Penrose (1989), Goldstein (2001), and Carroll (2010), who have recaptured and elaborated on Boltzmann's ideas. But I assume these authors will agree with me that the critical concepts and insights are all there in Boltzmann's original work and have stood the test of time. Reintroducing them to physicists, mathematicians, and philosophers proved to be highly necessary, however, as they were – and still are – subject to both unnecessary misunderstanding and important efforts to explore their full potential.

9.1 The Second Law of Thermodynamics

Our discussion is concerned with the explanation of the irreversible thermodynamic behavior of macroscopic systems. The term "thermodynamic behavior" refers to the ubiquitous phenomenon that physical systems, prepared or created in a non-equilibrium state and then suitably isolated from the environment, tend to evolve to and then stay in a distinguished macroscopic configuration called the *equilibrium state*. Familiar examples are the expansion of a gas, the dissipation of heat, the mixing of milk and coffee, and so on.

Phenomenologically, this empirical regularity is captured by the *second law of ther-*

modynamics positing the monotonous increase of a macroscopic variable of state called *entropy* which attains its maximum value in equilibrium.[1] One of the main tasks of *statistical mechanics* is to explain this macroscopic regularity on the basis of the more fundamental laws guiding the behavior of the system's micro-constituents.

A key concept of Boltzmann's statistical mechanics is the distinction between microstates and macrostates. Whereas the microstate $X(t)$ of a system is given by the complete specification of its microscopic degrees of freedom, the macrostate $M(t)$ is specified in terms of physical variables that characterize the system on macroscopic scales (like its volume, pressure, temperature, and so on). The macroscopic state of a system is completely determined by its microscopic configuration, that is $M(t) = M(X(t))$, but one and the same macrostate can be realized by a large (in general infinite) number of different microstates, all of which "look macroscopically the same." The partitioning of the set of microstates into sets corresponding to different macrostates is therefore called a *coarse-graining.*

Turning to the phase space picture of classical Hamiltonian mechanics for an N-particle system, a microstate corresponds to one point $X = (q, p)$ in phase space $\Gamma \cong \mathbb{R}^{3N} \times \mathbb{R}^{3N}$, $q = (q_1, q_2, ..., q_N)$ being the position and $p = (p_1, p_2, ..., p_N)$ the momentum-coordinates of the N particles, whereas a macrostate M corresponds to an entire region $\Gamma_M \subseteq \Gamma$ of phase space, namely the set of all microstates that realize M. The microscopic laws of motion are such that any initial microstate X_0 determines the complete micro-evolution $X(t) = \phi_t(X_0)$ of the system – represented by a unique trajectory in phase space going through X_0 – thereby also determining the macro-evolution $M(X(t))$ as the microstate passes through different macro-regions.

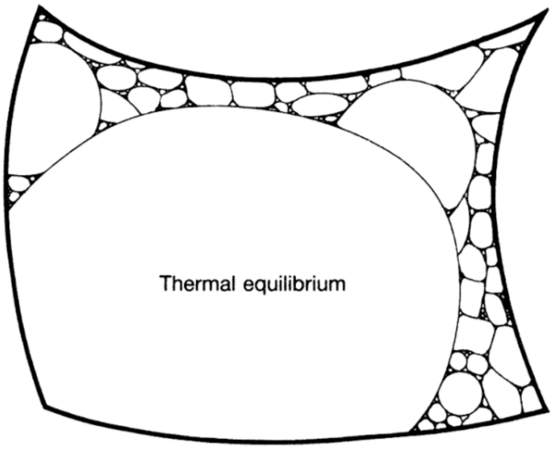

Figure 9.1: Partition of phase space into macro-regions. Size difference are vastly larger than depicted. Graphic from: Penrose (1989).

These concepts are pretty much forced on us if we accept the supervenience of macroscopic facts on microscopic facts, and they are essential to appreciating the

[1] This broad use of the term "second law of thermodynamics" may be somewhat ahistoric but has become customary in physics and I shall take license to adopt it, as well.

problem at hand. The second law of thermodynamics describes an empirical regularity about the *macro-evolution* $M(t)$ of a physical system. This macro-evolution is determined by the evolution of the microscopic configuration, which follows exact deterministic laws of motion. The aspiration of statistical mechanics is thus to explain or justify the empirical regularity expressed by a macroscopic law on the basis of the more fundamental microscopic theory. When it comes to the second law of thermodynamics, this seems like a quite formidable task, as it requires us to reconcile the *irreversibility* of thermodynamic behavior with the *time-reversal symmetry* of the microscopic laws of motion. The task was nevertheless accomplished by Ludwig Boltzmann in the second half of the 19th century. His account is based on two essential insights:

1. The identification of the Clausius entropy of thermodynamics with the (logarithm of the) phase space volume corresponding to a system's current macrostate. Formally:

   $$S = k_B \ln |\Gamma_{M(X)}|, \tag{9.1}$$

   where k_B is the Boltzmann constant and $|\Gamma_M|$ denotes the volume (the Lebesgue or Liouville measure) of the phase space region Γ_M, comprising all microstates $X \in \Gamma$ that realize the macrostate M.

2. The understanding that the *separation of scales* between the microscopic and the macroscopic level leads to enormous differences in the phase space volume corresponding to states with different values of entropy. In particular, we will generally find that the equilibrium region – by definition, the region of maximum entropy – is vastly larger than any other macro-region, so large, in fact, that it exhausts almost the entire phase space volume (or more precisely, at fixed total energy E, the induced $6N-1$-dimensional volume on the energy surface Γ_E). In other words: nearly every microstate *is* an equilibrium state.

These two points are related as follows: Entropy is an *extensive* state variable, meaning that, for fixed values of the other macro-variables, it usually grows like N, the number of microscopic constituents. Substantial entropy differences are thus of order N, and the differences in the measure of the corresponding macro-regions even of order $\exp(N)$. If we now recall that $N \sim 10^{24}$ for macroscopic systems (from Avogadro's constant), we see that the differences in phase space volume corresponding to different entropy levels are enormous. In other words, we generally find that for macroscopic systems, i.e., for systems with a very large number of microscopic degrees of freedom, the partitioning of microstates into macrostates does not correspond to a partitioning of phase space into regions of roughly the same size, but into regions whose sizes vary by a great many orders of magnitude, with the equilibrium region being by far the largest.

What we learn from these insights is, first and foremost, that the thermodynamic behavior we seek to explain is not a feature of certain *special* micro-evolutions, but rather the kind of macro-behavior that would obtain for almost any conceivable trajectory that a microstate, starting in a non-equilibrium region, could follow through phase space. In fact, the micro-dynamics would have to be very peculiar to *avoid* carrying the microstate into larger and larger phase space regions – corresponding to a gradual increase in entropy – and finally into the equilibrium region, where it remains for the foreseeable future. This is why Boltzmann's arguments are extremely robust against the details of the microscopic theory, giving us an understanding of thermodynamic behavior as a virtually universal feature of macroscopic systems.

Notably, though, it cannot be true that *all* microscopic initial conditions lead to an evolution of increasing (or non-decreasing) entropy. This is a straightforward consequence of the *time-reversal symmetry* of the microscopic laws, as was famously pointed out by Johann Loschmidt in his "reversibility objection." Hence, Lebowitz rightly warned us, quoting Ruelle, that Boltzmann's ideas are "at the same time simple and rather subtle" (Lebowitz, 1993b, p. 7). We will elaborate on these subtleties in the following section and see how the time-symmetry is broken.

The typicality account

To build on the basic principles of Boltzmann's statistical mechanics and go into the details of the typicality account, let us discuss the paradigmatic example of a gas in a box. We thus consider a system of $N \approx 10^{24}$ particles – interacting by a short-range potential (or not at all in the model of an *ideal gas*) – which are confined to a finite volume within a box with reflecting walls. Now assume that we find or prepare the gas in the macrostate M_2 sketched below (Fig. 9.2), that is, we consider a particle configuration that looks, macroscopically, like a gas filling out about half of the accessible volume. What kind of macroscopic evolution should we expect for this system?

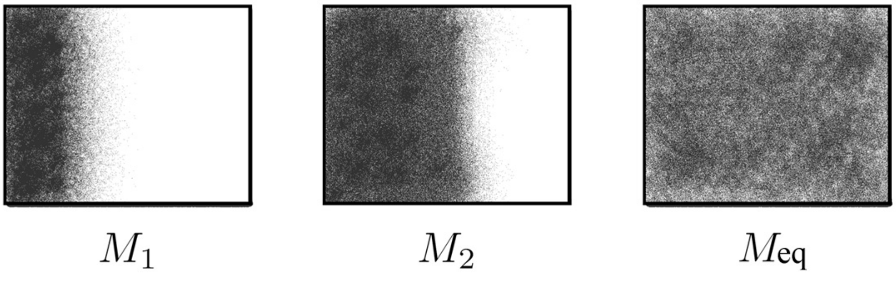

Figure 9.2: Thermodynamic evolution of an expanding gas

A simple combinatorial argument shows that *the overwhelming majority* of microstates that the system could possibly evolve into look, macroscopically, like M_{eq}, i.e., like a gas that is homogeneously distributed over the entire box. In fact, one can readily conclude that the phase space volume corresponding to this *equilibrium macrostate* M_{eq} is about $2^N \approx 10^{10^{24}}$ times (!) larger than the phase space volume occupied by configurations with substantially lower entropy (in agreement with our general reasoning above). Hence, as the particles move with different speeds in different directions, scattering from each other and occasionally from the walls, the system's microstate wanders around on an erratic path in the high-dimensional phase space, and we should expect that this path will soon end up in the equilibrium region $\Gamma_{M_{eq}}$ and then leave $\Gamma_{M_{eq}}$ only very rarely, corresponding to fluctuations of the entropy about its maximal value.

Larger fluctuations, e.g., from M_{eq} back into M_2, are possible, as well. They must, in fact, occur for almost all initial conditions according to Poincaré's *recurrence theorem.* However, as Boltzmann (1896a) already explained (see also Ehrenfest (1907)), the time-scales on which substantial fluctuations are to be expected are so astronomical – many orders of magnitude greater than the age of the universe – that they have no empirical relevance.

It was also clear to Boltzmann (at least after objections from Loschmidt) that there are initial conditions in Γ_{M_2} for which the system will not exhibit the "expected" thermodynamic behavior but follow an anti-thermodynamic trajectory of decreasing entropy. For if we consider a macrostate M_1 of even lower entropy, the time-reversal symmetry of the equations of motions implies that for every solution corresponding to a macro-evolution from M_1 to M_2, there exists another solution carrying an initial microstate in Γ_{M_2} into the lower-entropy macro-region Γ_{M_1}. (Indeed, we only have to take the solutions that have evolved from Γ_{M_1} into Γ_{M_2} and reverse all the particle momenta.) However, the microscopic initial conditions in Γ_{M_2} that lead to such an anti-thermodynamic evolution are *extremely special ones* relative to all possible microstates realizing M_2 – they form a subset whose measure relative to that of Γ_{M_2} is close to zero. The correct assertion is thus that *nearly all* initial configurations in Γ_{M_2} will evolve into the equilibrium-region $\Gamma_{M_{eq}}$, while only a very small set of "bad" initial conditions will show the anti-thermodynamic evolution from Γ_{M_2} into Γ_{M_1}. We will make these arguments more precise in a minute.

For now, let us note that it is more appropriate to consider not any individual trajectory, but the set of all solutions with initial conditions in Γ_{M_2}. The dynamics of a system of $N \approx 10^{24}$ particles is very chaotic, in the sense that even small variations in the initial configuration can lead to considerable differences in the time-evolution. Under the Hamiltonian dynamics, the set of microstates realizing M_2 at the initial time will thus quickly spread over phase space (respectively a submanifold compatible with the constants of motion) with the overwhelming majority of microstates ending up in the equilibrium-region and only a small fraction of "bad" initial configurations evolving into the comparably tiny macro-regions of equal or lower entropy.

Remark (On the concept of chaos)**.** The notion of "chaos" is difficult to exhaust with rigorous mathematical definitions. It is clear that some form of dynamical instability is characteristic of thermodynamic systems with many degrees of freedom, and there are various (usually highly idealized) mathematical concepts trying to capture this characteristic. Their fruitfulness in certain areas of mathematics may have contributed to the idea that one of them in particular must play a central role in the foundations of statistical mechanics and be identified as the precise dynamical assumption underlying Boltzmann's arguments. However, as emphasized before, the explanation of thermodynamic behavior is much more robust against the details of the microscopic model and doesn't hinge on any narrowly conceived property of the underlying dynamics. In particular, the relevant systems might easily fail to be ergodic, or mixing, or have everywhere positive Lyapunov exponents – to throw around some jargon – though their overall behavior would have to be completely qualitatively different from what it is generally understood to be in order to render the typicality account irrelevant.

Let us summarize our discussion up to this point. The time-reversal invariance of the microscopic laws implies that it cannot be true that *all* microstates in a low-entropy macro-region M_2 evolve into states of higher entropy. But Boltzmann's analysis tells us that micro-conditions leading to an entropy-decreasing evolution are *atypical* – they form a subset of extremely small measure – while *the great majority* of micro-configurations realizing M_2 will evolve into states of higher entropy until reaching equilibrium, and then stay in equilibrium for most of the time. Thermodynamic behavior, in other words, is *typical* given a non-equilibrium initial macro-state.

The role of the typicality measure

Throughout the above argument, the intuitive notions of *nearly all* or *extremely special*, that we associated with *typicality/atypicality*, were understood in terms of the stationary *Liouville measure*, i.e., in terms of the natural *phase space volume* of the set of microstates with the respective property. More precisely, for a perfectly isolated system with total energy E, we would have to consider instead of the Liouville measure the induced *microcanonical measure* μ_E on the hypersurface $\Gamma_E \subset \Omega$, to which the motion of the system is confined in virtue of energy conservation. For simplicity, I will often omit this distinction and merely refer to "phase space" and the "measure" or "size" of macro-regions.

In any case, a crucial property of the Liouville measure as well as the microcanonical measure is their *stationarity* under the microscopic time-evolution. This is such an essential feature because it means that

a) the notion of typicality is *timeless*, i.e., typicality statements do not have to make reference to external time.

b) the Hamiltonian dynamics "care" about the size of the macro-regions, in the sense that the stationary measure, as a measure on *initial conditions*, carries over to a well-defined measure on *solution trajectories*, which is such that the number of trajectories passing through a phase space region at any given time is proportional to the size of that region.

Turning back to Boltzmann's explanation of the second law, we note that the Liouville measure (respectively the microcanonical measure) as a *typicality measure* serves two purposes in the argument:

1. To establish that the region of phase space corresponding to the macrostate M_2 is very much larger than the region of phase space corresponding to the macrostate M_1, and that the region of phase space corresponding to the equilibrium macrostate M_{eq} is very much larger than the region of phase space corresponding to the macrostate M_2, so large, in fact, that it occupies almost the entire phase space.

   It is easy to learn about this "dominance of the equilibrium state" Frigg (2011), yet hard to appreciate the scale of proportions involved. Just think of the ratio $10^{10^{24}} : 1$ for the gas-model, which is beyond anything we could intuitively grasp.

   Together with the stationarity of the phase space measure, the dominance of the equilibrium state already implies that (by far) most of the solution trajectories are in equilibrium (by far) most of the time (see Reichert (2020) for a rigorous proof). This is not quite what we need, but not too far away, either.

2. To define a notion of typicality *relative to the current macrostate of the system*, allowing us to assert that *nearly all* microstates in the non-equilibrium region Γ_{M_2} will evolve into equilibrium, while *nearly all* equilibrium configurations will stay in equilibrium for the foreseeable future.

   Regarding the meaning of "nearly all," one should note that it is only in the idealized situation of a *thermodynamic limit* (where the number of microscopic degrees of freedom goes to infinity) that one can expect the exception set of "bad" configurations to be of measure *zero*, while if we argue about a realistic system, the *atypicality* of such configurations is substantiated by the fact that they have very very small (though positive) measure compared to the measure of the respective macro-region.

   In fact, stationarity allows us to estimate the measure of the good microstates relative to the bad ones in Γ_{M_2} by the ratio of phase space volume occupied by M_2 to the phase space volume corresponding to lower-entropy states. For let $\Gamma_{M_{low}}$ be the region of phase space corresponding to states of (substantially) lower entropy and let $B \subset \Gamma_{M_2}$ be the set of initial conditions in Γ_{M_2} that will have evolved into $\Gamma_{M_{low}}$ after a time Δt. Then $\Phi_{\Delta t}(B) \subseteq \Gamma_{M_{low}}$ and thus $|B| = |\Phi_{\Delta t}(B)| \leq |\Gamma_{M_{low}}|$, so that $|B| : |\Gamma_{M_2}| \leq |\Gamma_{M_{low}}| : |\Gamma_{M_2}| \approx 1 : 10^{10^{24}}$.

In sum, there are two typicality statements involved in Boltzmann's account of the second law. First, that equilibrium configurations are typical in Γ, i.e., with respect to all possible microstates. Second, that micro-configurations converging to equilibrium and thus leading to thermodynamic macro-behavior are typical relative to a non-equilibrium (initial) macro-region. The second statement is conceptually more subtle and much more difficult to prove rigorously.

9.2 Irreversibility

By incorporating into our discussion what was essentially Boltzmann's answer to Loschmidt's reversibility objection, we have already seen how the typicality account resolves the greatest challenge to our reductionist enterprise: the *prima facie* contradiction between the irreversibility of thermodynamic processes and the reversibility of the underlying mechanical laws. To highlight the solution, we recall that it was crucial to the typicality argument that it referred to (typical or atypical) initial conditions *relative to the initial macrostate.* Of course, in terms of overall phase space volume, a non-equilibrium macrostate occupies a vanishingly small fraction of phase space to begin with. The relevant notion of typicality when discussing convergence to equilibrium from a non-equilibrium macrostate M_2 is thus defined by the phase space measure conditioned on the fact that the initial microstate of the system is in the respective (low-entropy) region Γ_{M_2}.

Now, the time-symmetry of the microscopic laws is manifested in the fact that the phase space volume occupied by the bad initial conditions in Γ_{eq} – the initial conditions for which the system will evolve out of equilibrium into the macrostate M_2 – is just as large as the phase space volume occupied by the good initial conditions in Γ_{M_2} for which the system will relax into equilibrium. In other words, over any given period of time, there are just as many solutions that evolve *into* equilibrium, as there are solutions evolving *out* of equilibrium into a lower entropy state. The first case, however, is *typical* for systems in non-equilibrium, whereas the second case is *atypical* with respect to the possible equilibrium configurations.

It is this fact and this fact alone that establishes the irreversibility of thermodynamic behavior. And what breaks the time-symmetry is only the assumption (or preparation) of a special, i.e., low-entropy, initial macrostate.

Past Hypothesis and the thermodynamic arrow

By identifying the special macroscopic boundary conditions as the origin of the thermodynamic asymmetry, the typicality account is shifting the explanatory burden from why it is that a system in non-equilibrium relaxes to equilibrium (once macroscopic constraints are removed), to why it is that we find systems in such special states in the first place. Notably, with respect to *all possible microstates*, most configurations realize a state for which the system is in equilibrium, will be in equilibrium for most of its future, and has been in equilibrium for most of its past. This situation – which would be typical *simpliciter* – is indeed a time-symmetric one.

As long as we are preoccupied with boxes of gas or melting ice-cubes or the like, their low-entropy states will usually be attributable to influences from outside, i.e., to the fact that these systems are actually part of some larger system (usually containing a physicist, or a freezer, or the like) from which they "branched off" at some point to undergo a (more or less) autonomous evolution as (more or less) isolated subsystems. This presupposes, however, that these larger systems were *themselves* out of equilibrium; otherwise, they couldn't give rise to subsystems with less than maximal entropy without violating the second law.

If we think this through to the end, we arrive at the question why it is that we find *our universe* in such a special state, far away from thermodynamic equilibrium (and how to justify our belief that its state was even more special the farther we go back in the past). This is what Goldstein calls the "hard part of the problem [of irreversibility]" (Goldstein, 2001, p. 49), and it concerns, broadly speaking, the origin of irreversibility and the thermodynamic arrow of time in our universe. Dealing with the "hard problem" will require us to confront the meaning and status of the *Past Hypothesis* (Albert, 2000) postulating a very-low-entropy beginning of our universe. We will do so in Chapter 11.

9.3 The Status of Macroscopic Laws

In the old theory of thermodynamics, Clausius' second law

$$\frac{\mathrm{d}}{\mathrm{d}t}S \geq 0 \tag{9.2}$$

was, as the name says, understood as a law of nature. This understanding – together with still widespread skepticism about the atomic theory – was the major obstacle to appreciating the reduction achieved with Ludwig Boltzmann's statistical mechanics. Indeed, the Boltzmannian analysis forces us to make two concessions with respect to the nomological status of the second law of thermodynamics. First, it cannot hold *necessarily*, that is, for all possible initial micro-conditions. Second, it will not hold true *all the time* since on time scales approaching those of the Poincaré cycles, the Boltzmann entropy – the statistical mechanical counterpart of the Clausius entropy – *fluctuates*.[2] It will "only" be typically true on empirically relevant time scales.

Many contemporary publications are still putting a lot of emphasis on the fact that Boltzmann's second law is not exact, distinguishing, for instance, "thermodynamic-like behavior" – associated with the fluctuating Boltzmann entropy – from the "thermodynamic behavior" that was associated with the (supposedly) strictly non-decreasing Clausius entropy. In find this neither helpful nor necessary. Leaving aside the fact that these publications characterize "thermodynamic-like behavior" in terms of infinite-time averages that don't even capture the relevant phenomena, the basic point has been understood by physicists since one-and-a-half centuries ago, first and foremost

[2] Glenn Shafer has pointed out to me that it doesn't even make sense to say that a fluctuating quantity "increases" or "decreases" without a specification of the time scales to which such a statement refers.

by Boltzmann himself. Clausius' second law was not the full story, and the kind of "thermodynamic behavior" that some authors are still after is simply not in the cards.

Philosophically, the truly remarkable aspect about the statistical character of thermodynamic laws is not the way in which laws that were once thought of as exact turn out to be merely approximately true, but the way in which the regularities expressed by these laws turn out to be *contingent* rather than *necessary.* According to the microscopic theory, the initial condition of our universe could have been such that systems, prepared or created in a low-entropy state, would regularly end up on one of the "bad" trajectories that undergo an anti-thermodynamic evolution. That is to say that there are possible Newtonian universes in which gases are regularly found to contract rather than expand, in which heat does sometimes flow from colder to hotter bodies, and in which macroscopic objects such as apples and tables and chairs occasionally jump up in the air simply because a large number of particles in the ground happen to push in the same direction at the same time. In these possible universes, it is simply not true that such events are "very unlikely" because they happen all the time.

If we accept the microscopic laws as (more) fundamental, we thus have to concede that the so-called macroscopic laws – even in an approximate or statistical sense – are strictly speaking *not laws at all* in that they lack nomological necessity. And yet, I would insist, it is more than a mere contingency, more than a *factum brutum*, that thermodynamic regularities hold in our universe. The question we should ask is therefore: What more do the fundamental laws tell us about such regularities? What concept is weaker than necessity and captures the nomological status of the second law of thermodynamics?

The right answer, I submit, is not far to seek. Something is nomologically necessary if it obtains in *all* possible worlds permitted by the fundamental laws. Most "law-like" regularities fall a little short of that but obtain in *nearly all* possible worlds. The appropriate concept, in other words, is typicality.

The standard notions of probability are ill-suited to do the job. Epistemic probabilities are no substitute for nomological necessity. The fact that I do (or maybe *should*) believe that a macroscopic regularity obtains strikes me as categorically inadequate to bestow this regularity with any kind of nomological authority. Natural laws – including effective and special science ones – guide rational belief and expectations, but they are not themselves statements of (rational) belief. Frequentist probabilities are appropriate to characterize the *explanandum*, i.e., the regularity itself: over every short time-interval, the Boltzmann entropy *probably* increases. But they are question-begging when referring to a statistical distribution of initial conditions, and meaningless when referring to the universe as a whole.

If we are serious about our commitment to argue within the paradigm of a deterministic microscopic theory, we have to take it to the conclusion that there is nothing more random about the creation or preparation of a subsystem with certain initial conditions than about the evolution of an isolated system once prepared. To defer the source of "randomness" to the outside – from the box of gas to the shaky hands of the experimentalist, or to external perturbations preventing the subsystem from being perfectly isolated – is just to pass the buck. But the buck must stop, eventually, with the universe itself. For the universe is what it is; it exists once and only once, there is nothing before and nothing outside. And we either live in a universe that obeys the second law of thermodynamics (on cosmological scales and, with the possibility of very rare exceptions, in its branching subsystems) or we do not.

Typicality is just what the doctor ordered. It is a *modal* concept, expressing *objective* facts about the set of nomic possibilities referring, ultimately, to the universe as a whole. These facts provide for the relevant relation between the microscopic laws and the macroscopic regularity: the microscopic laws make the regularity typical. And they can ground *explanations* and *predictions*, thus serving the epistemic and behavior-guiding functions generally attributed to natural laws. Finally, typicality captures a sense of *counterfactual robustness* in that the regularity obtains not only for the actual initial micro-conditions of our universe but for nearly all possible ones.

Thermodynamic "laws" and other macroscopic phenomena are thus *typical regularities* under the fundamental microscopic laws. It is this fact that characterizes the relevant reduction and grounds their own law-like status.

On the derivation of typicality laws

In philosophy of science, the often-criticized yet very persistent models of Nagelian reduction and *deductive-nomological* explanations have established the idea that the relationship between a microscopic theory and a macroscopic regularity should be one of logical entailment. The macroscopic "law" or regularity should be derived from the more fundamental theory plus suitably specified "auxiliary assumptions." While this is not *entirely* wrong, understanding such a derivation in too naive logical terms misses the crucial role that *initial conditions* play in an account of a macroscopic phenomenon.

For what is it to *derive*, e.g., the thermodynamic behavior of a gas from the Newtonian laws of particle dynamics? Is it to show that there exists at least one microscopic configuration for which the gas will relax to equilibrium? Is it to show that it will happen for *all* possible initial states? The unsatisfactory weakness of the first proposition and the falsehood of the second must severely question the adequacy of purely deductive schemes of explanation.

Consider an inference of the form

$$\forall x (F(x) \Rightarrow G(x)),$$

where x ranges over possible (microscopic) realizations of the system and the predicate G is a suitable formulation of "exhibiting thermodynamic behavior." Then, the antecedent $F(x)$ would have to contain a clause that is essentially equivalent to the assumption "The initial conditions of the system x are such that $G(x)$," which makes the inference too trivial to be explanatory. *Of course* there exist initial conditions for which the gas will expand. There also exist initial conditions for which the gas will contract. And initial conditions for which the gas will transform into a banana. In other words, for a system with sufficiently many degrees of freedom and somewhat non-trivial dynamics, it is basically *always* possible to maintain that it has the (macroscopic) property G because the initial conditions were such that $G(x)$.

The only thing that can provide explanatory value in this context is the assertion of *typicality*, i.e., to establish that G is not a feature of certain special micro-conditions, but a physical fact that would obtain for *the great majority* of possible (initial) microstates. This is also to assure that the explanatory work is done as much as possible by the physical laws rather than some fine-tuned arrangement of microscopic degrees of freedom.

Notably, the relevant statement is now, logically and grammatically, a proposition about G rather than a proposition about any particular x. To provide an explanation or reduction of the second law of thermodynamics is thus not to state a set of assumptions about an *individual* system that implies its thermodynamic behavior, but to establish the relevant phenomenon as a typical feature of the microscopic laws.

Probabilistic schemes may give the appearance of a deductive explanation. If the explanandum is understood in probabilistic terms, it can be derived from the dynamical laws plus suitable probabilistic assumptions. We must, however, insist that any such account commit to a consistent interpretation of probability. Physical phenomena can often be described in *statistical* terms. But they cannot be derived from statistical (or frequentist) assumptions without "passing the buck," as argued before. Deductive probabilistic explanations – unless content with reducing one statistical regularity to another – are thus invariably playing some sort of "trick," like deriving physical phenomena from subjective belief, or smuggling in some extra-logical inference.

Without a resort to typicality, the weaker relation of *supervenience* faces the same problems as logical entailment. That a macroscopic regularity supervenes on the microscopic laws would mean that the regularity could not be different without a difference in the microscopic laws (or the relevant auxiliary assumptions). But this is patently false. For different micro-configurations, the same microscopic laws with the same macroscopic boundary conditions could give rise to completely different macro-behavior. Supervenience holds true only if the target of reduction is correctly understood as a typicality statement: There cannot be any difference in the *typical* regularities without a difference in the microscopic laws or the relevant macroscopic boundary conditions.

What else is left to say? Not much, I believe. To understand that a certain regularity is typical, and yet to wonder why it is that we observe this regularity in nature, is to wonder why our universe is typical, i.e., why it is, in the relevant respect, like the overwhelming majority of possible universes allowed by the fundamental laws of nature. And while I wouldn't know how to answer – except again with Einstein's bon mot that "God is subtle, but not malicious" – the very question strikes me as utterly uncompelling. Explanations have to end somewhere. If we can establish that a certain property is typical for a particular kind of system, this should remove any sense of mystery or puzzlement as to why we find such systems instantiating the respective property. Hence, we should consider the phenomenon to be conclusively *explained* on the basis of the microscopic theory. Similarly, if we can establish that a macroscopic feature or regularity is typical for a certain kind of system, we should expect to find this feature realized in systems of said kind. In this sense, it constitutes a *prediction* of the microscopic theory.

In this fashion, typicality statements figure in a way of reasoning about natural laws. In fact, since the situation in which we find ourselves towards the world is necessarily one in which all we can ever hope to know is compatible with a plurality of fundamental (microscopic) matters of fact, the relevant *explanatory* and *behavior guiding* statements that we can extract from fundamental physics are almost always results about typical solutions.

Typicality facts thus have certain *normative* implications, which shouldn't be confused for *logical* ones. Logically, the fact that a property G is typical doesn't entail anything about any *particular* instance (such an inference is simply not in the cards). It is always *possible* for a particular system – and ultimately our universe – to be atypical in the relevant respect. But facts that strike us as atypical are usually the kind of facts that cry out for *further* explanation. This is why a Casino manager has not just economic interest but reasonable grounds to suspect cheating if a player hits three jackpots in a single night. And this is why good scientific practice would eventually require us to revise our theory and look for a different set of laws, rather than endorsing an explanation of phenomena based on special initial conditions or, if you wish, a streak of bad luck. In the end, it is not logically but epistemically inconsistent to accept a physical theory and, at the same time, that our universe is (in the relevant respects) an atypical model of that theory, for this would undermine any reasons to endorse the theory in the first place.

9.4 Intermezzo: Time-Reversal invariance

I have repeatedly emphasized the time-reversal invariance of the microscopic laws multiple times without ever making precise what that symmetry is. My impression is that the concept of time-reversal is quite uncontroversial in physics, though not so among philosophers. Hence, some remarks (going beyond Newtonian mechanics) may be in order.

Newtonian Mechanics

Consider an N particle Newtonian system following an equation of motion

$$m\ddot{X}(t) = F\left(X(t)\right), \tag{9.3}$$

where $X(t) = (\boldsymbol{x}_1(t), ..., \boldsymbol{x}_N(t)) \in \mathbb{R}^{3N}$ is the (spatial) configuration of the system at time t, $m = \operatorname{diag}(m_1, \ldots, m_N)$ is the mass matrix, and F an (N-particle) force field. Let $X(t),\ t \in [0, T]$ be a solution of (9.3). We call

$$\bar{X}(t) := X(T - t),\ t \in [0, T] \tag{9.4}$$

the *time-reversal* of $X(t)$ and the evolution of $X(t)$ *reversible* if its time-reversal $\bar{X}(t)$ is also a solution of (9.3), i.e., a possible Newtonian history. Simply put: whatever can unfold in one time-direction can also happen in reverse.

Now, by taking the time-derivatives,

$$\begin{aligned} \dot{\bar{X}}(t) &= \frac{\mathrm{d}}{\mathrm{d}t} X(T - t) = -\dot{X}(T - t) \\ \ddot{\bar{X}}(t) &= \frac{\mathrm{d}^2}{\mathrm{d}t^2} X(T - t) = \ddot{X}(T - t), \end{aligned} \tag{9.5}$$

and thus, using the fact that $X(t)$ is a solution of (9.3), we have

$$m\ddot{\bar{X}}(t) = m\ddot{X}(T - t) = F\left(X(T - t)\right) = F(\bar{X}(t)). \tag{9.6}$$

Hence, if $X(t)$ is a solution of the Newtonian law (9.3), its time-reversal $\bar{X}(t)$ is also a solution. In other words, all Newtonian evolutions are reversible and we therefore call the laws of the form (9.3) *time-reversal invariant.*

Remark (On time-reversal)**.**

1. Since the law is also time-translation invariant, it doesn't matter if we consider $\bar{X}(t) := X(T - t)$ or $\bar{X}(t) := X(-t)$ as the time-reversal.

2. Except for special cases, e.g., $X(t) \equiv$ *const.*, $\bar{X}(t)$ and $X(t)$ are (mathematically) different trajectories in configuration space. Time-reversal invariance does not mean that *individual* solutions are time-symmetric but that the time-reversal of any solution is again a solution of the dynamical laws.

3. The time-reversal of $X(t)$ goes through the same *spatial* configurations in opposite order, while the *velocities* are reversed.

4. Time-reversal invariance would not hold in the presence of *dissipative* forces $F = F(X, \dot{X})$; but such Newtonian force laws are generally considered to be effective (e.g., friction) rather than fundamental.

Electrodynamics

For simplicity, we shall not consider the full Maxwell-Lorentz theory but only a single charge in an external electromagnetic field $(\boldsymbol{E}(t, \boldsymbol{x}), \boldsymbol{B}(t, \boldsymbol{x}))$. The Lorentz force law then reads

$$m\ddot{\boldsymbol{x}}(t) = q\left[\boldsymbol{E}\left(t, \boldsymbol{x}(t)\right) + \dot{\boldsymbol{x}}(t) \times \boldsymbol{B}\left(t, \boldsymbol{x}(t)\right)\right]. \tag{9.7}$$

Let $\boldsymbol{x}(t), t \in [0,T]$ be a solution of (9.7) and $\bar{\boldsymbol{x}}(t) := \boldsymbol{x}(T-t)$ its time-reversal. Analogously, the "naive" time-reversal of the electromagnetic field would be $(\bar{\boldsymbol{E}}(t), \bar{\boldsymbol{B}}(t)) := (\boldsymbol{E}(T-t), \boldsymbol{B}(T-t))$. In view of (9.5), it is easy to check that $\bar{\boldsymbol{x}}(t)$ satisfies

$$m\ddot{\bar{\boldsymbol{x}}}(t) = q\left[\bar{\boldsymbol{E}}\left(t, \bar{\boldsymbol{x}}(t)\right) - \dot{\bar{\boldsymbol{x}}}(t) \times \bar{\boldsymbol{B}}\left(t, \bar{\boldsymbol{x}}(t)\right)\right]. \tag{9.8}$$

But this is *not* the Lorentz force law (9.7) for the time-reversed fields $(\bar{\boldsymbol{E}}(t), \bar{\boldsymbol{B}}(t))$ (note the minus sign in front of the velocity-dependent term). However, $\bar{\boldsymbol{x}}(t)$ is a solution of (9.7) for the electromagnetic field

$$\left(\widetilde{\boldsymbol{E}}(t), \widetilde{\boldsymbol{B}}(t)\right) := \left(\boldsymbol{E}(T-t), -\boldsymbol{B}(T-t)\right), \tag{9.9}$$

i.e., we have:

$$m\ddot{\bar{\boldsymbol{x}}}(t) = q\left[\widetilde{\boldsymbol{E}}\left(t, \bar{\boldsymbol{x}}(t)\right) + \dot{\bar{\boldsymbol{x}}}(t) \times \widetilde{\boldsymbol{B}}\left(t, \bar{\boldsymbol{x}}(t)\right)\right]. \tag{9.10}$$

Indeed, the transformation (9.9) is also consistent with time-reversal in the Maxwell field equations, omitted in our discussion.[3]

Therefore, most people – with the notable exception of David Albert (2000) – agree that classical electrodynamics is time-reversal invariant, but that the fields transform non-trivially under time-reversal, namely into $(\widetilde{\boldsymbol{E}}, \widetilde{\boldsymbol{B}})$ rather than $(\bar{\boldsymbol{E}}, \bar{\boldsymbol{B}})$.

Albert (2000) argues that classical electrodynamics is not time-reversal invariant because the fields are part of the physical state and therefore any solution $(\boldsymbol{x}(t), \boldsymbol{E}(t), \boldsymbol{B}(t))$ would have to be reversible in the sense that $(\bar{\boldsymbol{x}}(t), \bar{\boldsymbol{E}}(t), \bar{\boldsymbol{B}}(t))$ is again a solution of the (Maxwell-)Lorentz equations. The mainstream view considers it sufficient that the particle-evolution $\boldsymbol{x}(t)$ is reversible, while the fields transform in a simple, albeit non-trivial way. The debate is thus over which parts of the physical state evolution have to be "exactly" reversible.

Quantum mechanics

The fundamental dynamical equation in quantum mechanics is the Schrödinger equation

$$i\hbar\,\partial_t\psi = H\psi, \tag{9.11}$$

describing the time-evolution of the wave function ψ. Given a solution $\psi(t), t \in [0,T]$ of (9.11), we consider the naive time-reversal $\bar{\psi}(t) := \psi(T-t)$, $t \in [0,T]$. It is now easy to check that $\bar{\psi}(t)$ satisfies

$$-\,i\hbar\,\partial_t\bar{\psi}(t) = H\bar{\psi}(t). \tag{9.12}$$

Again, this is not the correct law, but off by a minus sign. However, taking the complex conjugate of (9.12) (noting the imaginary unit on the left-hand-side and that

[3] Although it would not be hard to see from the inhomogeneous Maxwell equations $\nabla \times \boldsymbol{E} = -\partial_t \boldsymbol{B}$ and $\nabla \times \boldsymbol{B} = \boldsymbol{B} + \frac{1}{c^2}\partial_t \boldsymbol{E}$ why the $\boldsymbol{B}$-field picks up a minus sign under a reparametrization $t \to T-t$.

the Hamiltonian is real), we get:

$$i\,\hbar\,\partial_t \bar{\psi}^*(t) = H\bar{\psi}^*(t). \tag{9.13}$$

In other words, the naive time-reversal $\bar{\psi}(t) = \psi(T-t)$ does not solve the Schrödinger equation, but its complex conjugate

$$\widetilde{\psi}(t) := \psi^*(T-t) \tag{9.14}$$

does. This corresponds to an anti-unitary time-reversal operator, if one wants to be fancy about it.[4]

So why does this make quantum mechanics time-reversal invariant?[5] The orthodox response is that the empirically relevant quantity is not $\psi(t)$ itself but the probability density $|\psi(t)|^2$, or maybe expectation values $\langle\psi(t)|\hat{A}\,|\psi(t)\rangle$ of self-adjoint operators. The first is evidently the same for ψ and ψ^*, the latter remain the same up to a sign (the expected momentum, for instance, gets reversed, as it should). However, for the non-positivist, there would seem to be a difference between an invariance of the observable quantities and a fundamental symmetry of nature. The orthodox view is thus somewhat unsatisfying.

Bohmian mechanics

As usual, the situation is clearer in Bohmian mechanics. Here, what needs to be reversible is again the evolution of the particle configuration $X(t) = (x_1(t), ..., x_N(t)) \in \mathbb{R}^{3N}$ following the guiding equation

$$\dot{X}(t) = v^{\psi}(X(t)) = \frac{\hbar}{m}\mathrm{Im}\frac{\psi^*\nabla\psi}{\psi^*\psi}(X(t)). \tag{9.15}$$

Then, given a solution $X(t)$, $t \in [0,T]$ of (9.15), it is easy to check that its time-reversal $\bar{X}(t)$ solves

$$\dot{\bar{X}}(t) = \mathbf{v}^{\tilde{\psi}}(\dot{\bar{X}}(t)) \tag{9.16}$$

i.e., the guiding equation for the time-reversed wave function (9.14). Therefore, Bohmian mechanics has time-reversal symmetry.

The situation here is pretty much analogous to that in classical electrodynamics: The evolution of the particle configuration is reversible, but there are additional degrees of freedom – here the wave function, there the fields – that must transform in a non-

[4] López (2019) suggests instead a unitary time-reversal T such that $T^*HT = -H$. This doesn't exist for simple mathematical reasons. The spectrum of Hamiltonians is usually bounded from below, i.e., $\sigma(H) = [\beta, +\infty)$ and thus $\sigma(-H) = [-\infty, -\beta)$, but any unitary transformation $H \to T^*HT$ leaves the spectrum invariant.

[5] In quantum field theory, more precisely the standard model of particle physics, time-reversal invariance is violated in weak interactions, though the combined symmetry CPT (charge, parity, and time) is still considered a fundamental – and necessary – symmetry of nature, see, e.g., Streater and Wightman (2000).

trivially way.

One metaphysical view distinguishes the *primitive ontology* (PO) – the fundamental constituents of matter postulated by a theory – from degrees of freedom that belong to the *dynamical* or *nomological structure* of theory, whose role, in other words, is first and foremost a dynamical one for the evolution of the PO. Under symmetries in general, and time-reversal, in particular, the history of the PO has to be invariant (or covariant in the natural way), while the dynamical quantities can transform non-trivially. Elsewhere, I have endorsed the position that in Bohmian mechanics and classical electrodynamics the particles alone are the primitive ontology while the wave function, respectively the electromagnetic field, falls into the latter category (Esfeld et al., 2014; Lazarovici, 2018). In any case, the debate about which quantities should be exactly reversible – and thus the debate about which theories have bona fide time-reversal symmetry – is (where not based on mere misunderstandings) a debate about physical ontology.

Time-reversal symmetry and the arrow of time

Finally, there is the argument from time-reversal invariance against a primitive (metaphysical) direction of time. Some vague version of the argument seems to be almost folklore among physicists while some philosophers are still debating what the argument is supposed to be exactly. Although it is beyond the scope of our current discussion – which deals with the thermodynamic arrow of time – let me briefly state how I understand it.

Suppose there was a primitive direction of time. Then, any possible world, that is, any solution of the fundamental dynamical laws, would have a time-reversed “twin” which is physically identical except for the order in which the (particle) configurations are run through. At least, the two worlds would be indistinguishable by subjects inhabiting them, assuming that subjective experience supervenes on the spatio-temporal distribution of matter. Thus, by the principle of the identity of indiscernibles (PII) – or, more profanely, a principle of parsimony – the twin solutions should be regarded merely as different mathematical representations of one and the same physical world.

In particular, we represent the time axis by the continuum of real numbers. But in fact, $\mathbb{R}$ has mathematical surplus structure in that it is totally ordered, and reversing the orientation – by the reparameterization $X(t) \to X(-t)$, is merely a different choice of gauge that doesn't correspond to any difference in the world. Consequently, time-reversal must be understood as a *passive* transformation since the idea that one could (hypothetically) hold the direction of time fixed while changing the order of events unfolding in it is meaningless. In other words, the structure of time involves only a betweenness relation (a triadic relation between points in time) but not a dyadic order relation of “earlier than” and “later than.”[6]

[6] This is sometimes called a “C-theory” of time, see, e.g., Farr (2018).

I find this view attractive but must admit that the argument from time-reversal symmetry is not hard to dismiss. One can simply reject the PII or deny that the parsimony of the deflationary view has any virtue. One can insist that subjective experience supervenes not only on a world's unoriented path in configuration space but also on the direction of change or "passage." Maybe more convincingly, one can insist that a solution trajectory and its time-reversed twin are not physically identical, even in the Newtonian case, since the velocities/momenta differ by a global sign. At this point, both sides of the debate would seem to beg the question because a reversal of all particle velocities amounts to a physical difference if and only if there is a direction of physical time itself.

9.5 Objections and Responses

In this final section, we are going to address the most common objections to the typicality account that have been raised in the contemporary philosophical literature. We will argue that these objections are unfounded and often based on unnecessary misunderstandings of Boltzmann's arguments.

Missing the point of typicality

One of the most common mistakes in the debate about Boltzmann's statistical mechanics is the failure to appreciate the difference between a typicality statement and an inference about particular instances. Consider, for example, the objection of Roman Frigg in reply to Goldstein (2001):

> Goldstein suggests that a system approaches equilibrium simply because the overwhelming majority of states in Γ_E are equilibrium microstates [...]. This is wrong. If a system is in an atypical microstate [...], it does not evolve into an equilibrium microstate just because the latter are typical; typical states do not automatically function as attractors. (Uffink, 2007, 979–980) provides the following example. Consider a trajectory $x(t)$, i.e., the set $\{x(t) = \phi_t(x(t_0)) \mid t \in [t_0, \infty)\}$, a set of measure zero in Γ_E. Its complement, the set $\Gamma_E \backslash x(t)$ of points not laying on $x(t)$, has measure one. Hence the points on $x(t)$ are atypical while the ones not on $x(t)$ are typical (with respect to Γ_E, μ, and the property 'being on $x(t)$'). But from this we cannot conclude that a point on $x(t)$ eventually has to move away from $x(t)$ and end up in $\Gamma \backslash x(t)$; in fact the uniqueness theorem for solutions tells us that it does not. The moral is that non-equilibrium states do not evolve into equilibrium states simply because there are overwhelmingly more of the latter than of the former, i.e., because the former are atypical and the latter are typical. (Frigg, 2011, p. 82).

Of course, no one suggests, in the naive sense implied by Frigg, that any *particular* trajectory must move to equilibrium “simply because” the overwhelming majority of states are equilibrium states, just as no one suggests that any *particular* lottery ticket must lose simply because the overwhelming majority of possible combinations are losing combinations. Goldstein’s argument – which is the same as our argument, which is the same as Boltzmann’s – is not about what any individual solution trajectory must do, but about what the great majority of them does.

So what is the point of the “counterexample” formulated by Uffink that made such an impression on Frigg? Evidently, it is correct that a solution $x(t)$ will never enter the phase space region $\Gamma_E \setminus \{x(t)\}$ despite the fact that this set has measure one. But *almost all solutions* will. In fact, it follows from the “uniqueness theorem” that every other solution (with the same total energy) lies entirely in $\Gamma_E \setminus \{x(t)\}$. Quite possible, the critics have misunderstood what the really interesting typicality statement in the proposed account is: not that equilibrium is typical, but that convergence to equilibrium is typical relative to any non-equilibrium initial macrostate.

Hence, leaving aside the fact that the artificial phase space region considered by Uffink is of no physical interest, the alleged “counterexample” is completely off-target. It’s like some people arguing that typical lottery tickets will fail to win the jackpot because of the huge number of possible combinations, while others are running around with a winning ticket in order to disprove them.

The measure zero problem

If Uffink’s example works at all, then as another instance of the so-called *measure zero problem* which, in a nutshell, is the observation that as soon as we go to a more fine-grained description, any physical system is found to be atypical with respect to some properties (see our discussion in 1.2). In particular, for a continuous state-space and a nonsingular measure, the *actual* microscopic configuration and, as just noted, even the entire trajectory of a system will constitute a set of measure zero. While this observation is often presented as a serious challenge to typicality arguments (see, e.g., Sklar (1973)) I don’t see it as causing much of an embarrassment for the way of reasoning defended here.

There are facts and regularities that are explained by the microscopic laws by virtue of being typical (like ice-cubes melting at high temperature). There are contingent facts about physical systems that are not typical but can be explained in a different way, usually by tracing them back to other special states of affairs. For instance, the current state of my desk is certainly atypical with respect to the exact distribution of objects on it, but I could tell some sort of causal story about how a used coffee mug ended up near the keyboard and how the blue book came to lie on top of the heavier red one. Finally, there are facts like the one that a trajectory through some state space will never cross its complement – which do not require any explanation but can be used to create unnecessary confusion.

That said, there is a serious question about what, in general, characterizes a good explanandum; why some measure zero events would de facto falsify our theory while others are reasonably accepted as brute facts. I provided my best attempt at an analysis in 1.2.

Misidentifying the macrostates

One objection to the typicality account seems to go back to Lavis (2005) and was later picked up by other authors, in particular Werndl and Frigg in several publications (Frigg and Werndl, 2012; Werndl and Frigg, 2015a,b). This objection questions one of Boltzmann's fundamental insights by arguing that the equilibrium macrostate (the state of maximal entropy) does not generally take up the majority of phase space volume. The corresponding macro-region may be the biggest one (in terms of the natural phase space measure) but not bigger than all the non-equilibrium regions combined.

Lavis' argument follows Boltzmann's combinatorial analysis of the *gas in a box*, in which the one-constituent state space is partitioned into finitely many cells while counting the number of particles in each cell. Lavis observes – considering, for instance, the simple case of $N = 8$ particles distributed over $m = 4$ cells – that while the most likely occupation $(2, 2, 2, 2)$ (meaning: every cell contains exactly two particles) corresponds to more phase space volume than, say, $(3, 2, 2, 1)$, there are 12 possible permutations of $(3, 2, 2, 1)$, all describing non-equilibrium macrostates. Hence, he continues, the sum of the measures of such degenerate states exceeds that of the largest "macrostate" $(2, 2, 2, 2)$ which Lavis takes to be the Boltzmann equilibrium.

I am surprised that this could have caught on as a serious objection. Of course, while having *exactly* N/m particles in each of the m cells is more likely, i.e., corresponding to larger phase space volume, than any other specific occupation of cells (assuming m even divides N), this is overall a very *special* configuration. In fact, the "probability" of this exact equidistribution goes to zero for large N. However, for large (macroscopic) N and small (microscopic) particles, having precisely N/m particles in each cell is *macroscopically indistinguishable* from configurations in which some cells contain one or two or ten or even a million particles more than others. In other words, the exact equidistribution (that Lavis falsely identifies with the Boltzmann equilibrium) and small deviations from the exact equidistribution (that he wants to weigh against the former) actually coarse-grain to *one and the same* macrostate, all corresponding to thermodynamic equilibrium.

More precisely, it is an elementary result of probability theory that the phase space measure is concentrated on configurations for which the number of particles in each cell deviates by at most $\sim \sqrt{\frac{N}{m}}$ from the mean value. For $N \approx 10^{24}$ and $m \ll N$, this means that microstates corresponding to local density-fluctuations of less than a billionth of a percent exhaust almost the entire phase space volume. It is this set of for all practical purposes indistinguishable configurations that constitute the relevant

Boltzmann equilibrium.

To suggest that a gas is “out of equilibrium” if there are two more molecules in the left half of the box (let’s say) is to miss the whole point of the micro/macro-distinction and to attack a caricature of Boltzmann’s ideas.

In a similar vein, Werndl and Frigg (2015b) criticize Penrose and Goldstein for falsely inferring the *dominance* of the equilibrium state (the equilibrium region exhausting a majority of phase space volume) from its *prevalence* (the equilibrium region being larger than any non-equilibrium region) “by calculating that the ratio between the measure of the equilibrium macro-region and the macro-region of a standard non-equilibrium state is of order 10^N.” The objection here is, again, that the measure of the non-equilibrium regions could still sum up to a total that exceeds the measure of the equilibrium region. But if we recall that the scale of N is roughly 10^{24} for macroscopic systems, one must wonder how many different macrostates Werndl and Frigg believe that one can meaningfully distinguish. Even if there were measurements fine enough to discern $10^{10^{24}}$ different states of a gas or a cup of coffee (which there aren’t), they would clearly cease to be “macroscopic” in a sense relevant to statistical mechanics.

In conclusion, the objection presented by these authors has little to do with a problematic “degeneracy” of lower-entropy states and everything to do with a failure to consider an appropriate coarse-graining into macrostates. While one cannot completely discard the possibility of peculiar counterexamples (and there are, indeed, cases in which it makes sense to speak of two or more equilibrium states), the dominance of the Boltzmann equilibrium is the generic case in statistical mechanics – when “Boltzmann equilibrium” is understood correctly.

The role of the dynamics

A final point of criticism that we must address is that Boltzmann’s account fails to identify precise assumptions about a system’s micro-dynamics that imply the typicality of thermodynamic behavior. Part of my response to this criticism can be found in my previous remarks about the elusive character of “chaos,” as well as in the observation that there is nothing special (in the colloquial sense) about dynamics carrying microstates from the vanishingly small non-equilibrium region into the rest of phase space corresponding to thermodynamic equilibrium. Another part of the response consists simply in the concession that the typicality account is not – and doesn’t pretend to be – a rigorous mathematical proof.

The real debate, however, goes somewhat deeper and concerns the nature of physical explanation and the role of mathematical proof in general. It deserves its own section.

9.6 Physics and Mathematics

> The intellectual attractiveness of a mathematical argument, as well as the considerable mental labor involved in following it, makes mathematics a powerful tool of intellectual prestidigitation – a glittering deception in which some are entrapped, and some, alas, entrappers.
>
> — Jack Schwartz, "The Pernicious Influence of Mathematics on Science" (1996)

Here is a joke: *How can you tell that your janitor studied philosophy of physics? You complain that the heat is not working, and he asks you to check if your living room has good ergodic properties.*

It is said that a joke is no good if one has to explain the punchline. But since I have limited comedic ambitions, I am going to do exactly that. The dispersion of heat in a volume, such as your living room, is a thermodynamic process, an instance of the *second law*, in fact. It can be phenomenologically described by the heat equation but is well-understood, from a more fundamental point of view, in terms of particle motion. The reduction of the phenomenological law to the micro-dynamics of particles falls into the domain of statistical mechanics. And it is a widespread belief in the philosophical (but also part of the physical) literature that the explanation of thermodynamic behavior – if not the success of statistical mechanics, in general, – rests on the assumption of *ergodicity*. If this were so, it would mean that physics does not provide good reasons to expect the dispersion of heat emitted from a radiator if the room – qua physical system – failed to be ergodic.

Now, one could be pedantic and point out that a living room is not a perfectly closed system, so it cannot be a proper ergodic system in the precise mathematical sense. And one could further point out that even if we modeled the living room as a closed ellipsoid with perfectly reflecting walls and the air molecules as little billiard balls bouncing of each other[7], ergodic properties will hardly survive the smallest departure from this idealization. Most importantly, however, ergodicity is such an abstract mathematical concept that the very question whether it applies to a living room strikes me as somewhat ridiculous, close to a category mistake.

Let us recall what ergodicity is all about. In the modern literature[8], ergodicity is usually introduced as a property of dynamical systems: A dynamical system is *ergodic* if every invariant set has measure 1 or 0. (A measurable subset $A \subseteq \Gamma$ is *invariant* under the time-evolution if $\phi_t(A) = A, \forall t$.) A solution, i.e., a flow-line, of the dynamical system may also be called ergodic, namely, if the proportion of time that the trajectory spends (over its entire history) in any region of phase space corresponds

[7] See Sinai (1970); Bunimovich (1979) for ergodic properties of analogous two-dimensional models.

[8] The "ergodic hypothesis" was first introduced by Ludwig Boltzmann but didn't even appear anymore in his second lectures on gas theory (1896). The concept was later revived, in modern form, by the groundbreaking works of Birkhoff, von Neumann, and Khinchin that established ergodic theory as a productive (and admittedly very elegant) field of mathematics, whose physical relevance, however, is questionable.

to the measure of that region. Formally:

$$\lim_{T\to\infty}\frac{1}{T}\int_0^T \mathbb{1}_A(X(t))\mathrm{dt} = \mu_E(A), \tag{9.17}$$

where $\mathbb{1}_A(x)$ is the characteristic function of $A \subseteq \Gamma_E$. This, in turn, is essentially equivalent to the statement that the solution-trajectory comes arbitrarily close to every single point in Γ_E, thus establishing the connection with Boltzmann's original (quasi)-ergodic hypothesis.[9] The celebrated ergodic theorem of Birkhoff (1931) establishes that typical solutions of an ergodic system – in the strong sense of *all except for a set of initial conditions with measure zero* – are ergodic trajectories.[10]

In the literature on foundations of statistical mechanics, ergodicity (or stronger properties higher up the "ergodic hierarchy") has been assigned various tasks: to justify the choice of the microcanonical measure as the unique stationary measure on the energy hypersurface, to explain the relevance of Gibbsian ensemble averages by identifying them with time averages of individual systems, or to account for thermodynamic behavior and the convergence to equilibrium. All of these ideas are misguided for different reasons, but for now, we shall focus on the latter claim that ergodicity is relevant for the microscopic reduction of the second law. Frigg and Werndl (2011) give the following argument:

> Consider an initial condition x that lies on an ergodic solution. The dynamics will carry x to [the equilibrium region] $\Gamma_{M_{eq}}$ and will keep it there most of the time. The system will move out of the equilibrium region every now and then and visit non-equilibrium states. Yet since these are small compared to $\Gamma_{M_{eq}}$, it will only spend a small fraction of time there. Hence the entropy is close to its maximum most of the time and fluctuates away from it only occasionally. Therefore, ergodic solutions behave [thermodynamic]-like. (p. 633)

In brief, ergodicity is not sufficient for thermodynamic behavior because

a) Ergodicity of trajectories is a time-symmetric property (the time-reversal of an ergodic solution is also an ergodic solution) and thus cannot account for thermodynamic irreversibility.

b) Infinite time averages imply nothing about the behavior of the system on empirically relevant time scales. The characteristic time scale associated with irreversible thermodynamic behavior is that of a system's *relaxation time* (the time it typically

[9] See, for instance, the Ehrenfests (1907) on Boltzmann's ergodic hypothesis or Sklar (1973).

[10] Frigg and Werndl (2011) advocate instead for a weaker notion of "epsilon-ergodicity" which only requires an ergodic evolution for all initial micro-conditions except for a set of positive measure $\leq \epsilon$. This is doing nothing to avoid our following objections; the uselessness of epsilon-ergodicity is only more obvious since non-equilibrium macro-region have tiny measure, to begin with, and may thus lie entirely in this exception set.

takes to reach equilibrium), which may be seconds for the spreading of a gas, minutes for the cooling of a hot bowl of soup, many years for the decay of radon, and many billions of years for the heat death of the universe. But all this is just the blink of an eye compared to the time scales associated with ergodic behavior (see Fig. 9.3). The time scales associated with ergodic behavior, the time scales, that is, on which trajectories begin to "wind around" the energy-hypersurface and explore even the smallest (macro-)regions, are those of the *Poincaré cycles* which were already estimated by Boltzmann, for the gas model, to be about $10^{10^{20}}$ years(!) – exceeding the age of our universe by many orders of magnitude. Conceptually, the proposition that a particular system behaves ergodically doesn't even make sense when referring to a period of time relevant to its thermodynamic evolution, just as the proposition that a particular gentleman is "aging in dignity" doesn't make sense when referring to a period of few nanoseconds.

And ergodicity is not necessary for thermodynamic behavior because

a') Given the huge differences in phase space volume corresponding to the equilibrium versus non-equilibrium regions, we do not need exact equality between phase and time averages to establish that a system will spend most of the time in equilibrium. To put it differently: virtually any conceivable path through phase space would spend most of the time outside the non-equilibrium region. To do so, a solution does not need to densely cover the entire phase space, any more than a person's travel route needs to cover the entire surface of the earth to account for her spending most of the time outside the Principality of Monaco.

b') We do not care if and for how long a micro-trajectory visits *every* measurable subset of phase space. Only the phase space regions associated with the partition into macrostates are relevant for describing a system's macro-evolution.

c') There are very instructive and well-studied toy models for convergence to equilibrium that are clearly not ergodic (see Bricmont (1995, 2001) for good discussions).

In effect, an ergodic evolution of a trajectory has nothing to do with thermodynamic behavior. Arnold and Avez, in their standard work on ergodic theory, put it even more concisely (1968, p. 77, footnote 17):

> *Statistical mechanics deals with asymptotic behavior as $N \to +\infty$ (N=number of particles) and not as $t \to +\infty$ for fixed N.*

There is certainly something about ergodicity as a property of *dynamical systems* that has the right flavor, in that it expresses a notion of chaos and implies the absence of dynamical "barriers" preventing solutions from reaching equilibrium (though invoking it for this purpose is, to adopt a German proverb, like shooting sparrows with a cannon). Overall, however, the idea that ergodicity plays a central role in the foundations

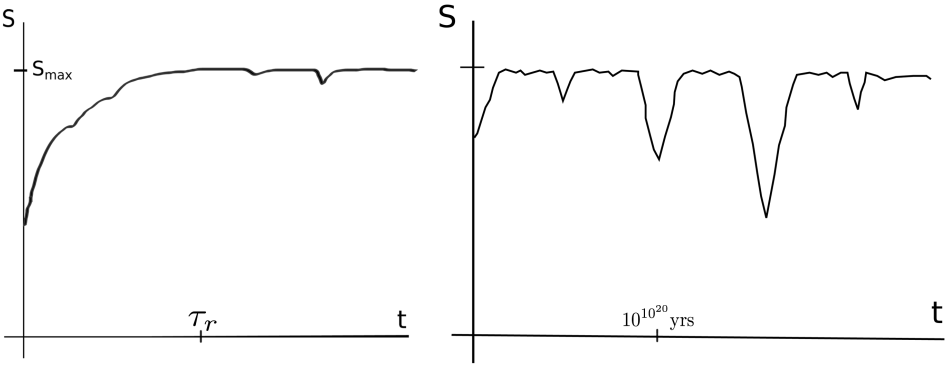

Figure 9.3: Typical entropy curves of macroscopic systems on thermodynamic time scales (left) and on ergodic time scales (right). τ_r is the relaxation time. On the right, periods of near-maximal entropy are vastly longer than depicted.

of statistical mechanics is misguided and has repeatedly lead the foundational debate astray.

I want to briefly discuss another ergodic property, which is stronger than ergodicity yet somewhat more instructive. A dynamical system is called *mixing* if

$$\lim_{t\to\infty} \lambda(A \cap \Phi_{-t}(B)) = \lambda(A)\lambda(B) \tag{9.18}$$

for all measurable $A, B \subseteq \Gamma$. We now consider a Hamiltonian dynamical system $(\Gamma, \lambda, \Phi_t)$ and a macro-variable F which is an average of one-particle quantities, i.e., a function of the form $F(x) = \frac{1}{N}\sum_{i=1}^{N} f(x_i),\ x = (x_1, ..., x_N) \in \Gamma$. The equilibrium region then corresponds to

$$B = \{x \in \Gamma : |\frac{1}{N}\sum_{i=1}^{N} f(x_i) - \bar{F}| < \epsilon\},\ \bar{F} = \int_\Gamma F(x)\mathrm{d}\lambda(x)$$

with $\epsilon \sim N^{-1/2}$. This is nothing more than the law of large numbers: micro-configurations for which the value of F deviates only slightly from the theoretical mean exhaust the great majority of phase space volume. We consider, however, systems that start out in a non-equilibrium macro-region A. This is to say that at any time $t > 0$, we care only about equilibrium configurations that have evolved from the macro-region A at $t = 0$. Such a boundary condition leads, in general, to correlations between the particles. But now we can use the mixing property, together with the (weak) law of large numbers, to conclude "convergence to equilibrium" in the following sense:

$$\lambda(\Phi_t(A) \cap B) = \lambda(A \cap \Phi_{-t}(B)) \xrightarrow{t\to\infty} \lambda(A)\lambda(B) \geq \lambda(A)\left(1 - \frac{\mathbb{V}(f)}{\epsilon^2 N}\right), \tag{9.19}$$

assuming the variance of f is finite. That is, in the limit $t \to \infty$, nearly all initial microstates in A end up in the equilibrium region B. This result (notably a typicality result) is as elegant as it is physically irrelevant. On the one hand, because realistic systems are hardly "mixing." On the other hand, because, yet again, nothing of empirical import follows from the infinite-time limit. This is in notable contrast to the quantitative estimate in terms of the particle number N on the right-hand-side. Simply put, the $N \to \infty$ limit coming from the LLN is physically relevant, while the $t \to \infty$ limit coming from the mixing property is – absent additional results about the convergence rate – pure mathematical abstraction. At the same time, the exact factorization in (9.19) is unnecessarily (and unrealistically) strong. What we would really need is $\lambda(A \cap \Phi_{-t}(B)) \approx \lambda(A)\lambda(B)$ (in an appropriate sense, which can be very weak) and on time scales that are long relative to the mean free time between particle collisions, but not much longer than the observed relaxation time of the system.

Here, we should keep in mind the insight that Pierre Duhem formulated more than a century ago:

> [A] mathematical deduction is of no use to the physicist so long as it is limited to asserting that a given rigorously true proposition has for its consequence the rigorous accuracy of some such other proposition. To be useful to the physicist, it must still be proved that the second proposition remains approximately exact when the first is only approximately true. (Duhem, 1954, p. 143)

Ergodic properties fare badly against this "principle of stability" (Fletcher, 2020), both in the sense that they appear to be unstable against variations of the microscopic model, and in the sense that they imply only exact but empirically irrelevant results.

Hence, just like the weaker notion of ergodicity, mixing is doing both too much and too little. And yet, like ergodicity, the mixing property captures – in a very Platonic sense – some idea of chaotic behavior that is both plausible and relevant for realistic macro-systems: After a certain number of scatterings, the particles "forget" their common origin in the non-equilibrium region A and acquire statistical independence. Consequently, the configuration will start to look more and more like a typical, i.e., equilibrium, configuration relative to the entire phase space. This is essentially what any rigorous derivation of thermodynamic behavior would have to show in one form or the other – though almost certainly not by establishing the mixing property. Instead, we would have to leave the Platonic realm of ergodic theory and get our hands dirty with very hard analysis and cumbersome epsilonics. Even for highly simplified models, a rigorous and physically relevant proof of convergence to equilibrium will look nothing like (9.19) or the one-paragraph argument quoted from Frigg and Werndl above.

Proof and explanation

The deeper moral here is that when it comes to the difficult problem of macro-to-micro reduction, mathematical physics is, in many ways, the art of the possible. Evidently, we cannot just solve the equations of motion for $N \approx 10^{24}$ particles to check for the desired macro-behavior. Hence, it lies in the nature of the problem that rigorous results are rare and difficult to come by. Instead, we make simplifications, approximations, and idealizations. We use cut-offs, rescalings, and infinite limits – alongside various formal assumptions that allow us to derive relevant estimates or satisfy the conditions of previously established theorems. And often, such assumptions will take precedence in the statement of a mathematical result while the critical ideas behind the strategy of proof get lost in technical details.

In consequence, not all proofs are explanatory, and not all explanations can be turned into rigorous proof. Discerning mathematical abstractions and technical crutches from physical insights that do actual explanatory work is a very subtle issue. When we discuss the philosophical foundations of statistical mechanics, it is tempting to look at mathematical publications and read the premises of the reported results as the relevant axioms or auxiliary assumptions for some sort of deductive-nomological explanation – especially when they come in such a simple and elegant form as ergodic properties do. This literal-mindedness about mathematics is counterproductive, however, leading us further away from a true understanding of the phenomena.

Most mathematical theorems in statistical mechanics are extremely valuabe by refining, substantiating, or challenging our physical understanding. But insisting on mathematical rigor is not always testament to a rigorous mind. Hand-waving about the fundamental postulates of a theory should not be acceptable (cf. the our discussion of the quantum measurement problem in Ch. 12); but when it comes to applying a theory in complex situations, arguments based on an educated intuition can be more instructive than precise yet sterile proof. For while it lies in the nature of a logical deduction that the truth of the conclusion depends rigidly on the truth of the premises, it is essential to a good physical explanation to be reasonably stable under perturbations of its underlying assumptions – in particular when they are themselves the result of approximations and idealizations (cf. Schwartz (1966)).

Boltzmann's account of thermodynamic irreversibility is an explanation or explanatory scheme, not a proof. It leaves many details to be filled out in individual cases, but its generality and robustness is what makes it so powerful and compelling. In the philosophical literature, the account has nonetheless come under attack for its lack of mathematical rigor and the alleged failure to make its assumptions about the micro-dynamics explicit (Uffink, 2007; Frigg, 2009, 2011; Frigg and Werndl, 2011, 2012). Frigg and Werndl (2012) even go as far as declaring that the typicality account is "mysterious" because the "connection with the dynamics" is unclear (p. 918). Jos Uffink writes on a similar note (as a conclusion to his "counterexample" discussed in the previous section):

> [I]n order to obtain any satisfactory argument why the system should tend to evolve from non-equilibrium states to the equilibrium state, we should make some assumptions about its dynamics. In any case, judgments like 'reasonable' or 'ridiculous' remain partly a matter of taste. The reversibility

> objection is a request for mathematical proof (which, as the saying goes, is something that even convinces an unreasonable person). (2007, p. 61)

We have already seen that these objections are at least partially based on a misunderstanding of what the typicality account actually argues for. That aside, the critics appear to insist that any satisfactory account of the second law must involve a precise mathematical assumption about a system's micro-dynamics that *logically implies* its thermodynamic behavior (see also (Frigg and Werndl, 2011, p. 632)). This request strikes me as overly ambitious and I have tried to explain why a "reasonable person" will often settle for less than rigorous proof. The promise of ergodic programs old and new was that the dynamics of a trillion trillion interacting particles can be abstracted to a simple mathematical property that is both precise and universal, i.e., realized by a great variety of relevant systems. I would be elated if such a property existed but see no reason why it should. When dealing with complex phenomena, precision usually comes with specificity, while the explanation provided by Boltzmann operates on a much more general level, thereby capturing a nearly universal truth.

That said, one of the great insights from Boltzmann's analysis is precisely that thermodynamic behavior does not rely on any special feature of the microscopic time-evolution. Simply put, the role of the dynamics is to carry a great majority of the microstates in the vanishingly small non-equilibrium region reasonably quickly into the rest of phase space that corresponds to thermodynamic equilibrium. And this is so much weaker and so much more plausible as an "assumption" about the micro-dynamics of complex systems that it is hard to see how it could be further explained by reducing it a formal mathematical premise.

If you throw a bottle into the Atlantic Ocean, what precise feature of oceanic currents ensures that it will typically spend most of the time outside the region where the Titanic sank?

Indeed, it is the absence of thermodynamic behavior that would point to some remarkable feature of the phase space flow (e.g., dynamical attractors) in need of more detailed investigation. I actually agree with Frigg and Werndl (2011) that the ideal result, from a technical point of view, would be yet another typicality statement: that typical Hamiltonians, within a relevant class of interacting models, lead to thermodynamic behavior and convergence to equilibrium. But at the current stage of mathematical research, I doubt that such a proof is in the cards, and I don't believe that our physical understanding of the second law of thermodynamics hinges on it in any significant way.

9.7 *H*-theorem and Kinetic Equations

Although the formula engraved on Boltzmann's tombstone is equation (9.1), connecting the entropy of a microstate with the measure of the corresponding macrostate, his name is at least as intimately associated with the *Boltzmann equation* and the *H*-

theorem, describing, in a more quantitative manner, convergence to equilibrium for a low-density gas. This H-theorem is of great interest in light of our previous discussion. First, because it illustrates very clearly the need for a typicality argument. Second, because it can be viewed as a concrete implementation of the general scheme that we introduced as the "typicality account."

By expanding on these points, I also want to counter two widespread misconceptions that may have arisen from Boltzmann's first presentation of the H-theorem but persisted despite his more refined argumentation in later writings. The first is manifested in the charge that the account of thermodynamic irreversibility provided the H-theorem begs the question because the derivation of the Boltzmann equation is based on an explicitly time-asymmetric assumption about the micro-dynamics. The second, more basic misunderstanding is that the H-theorem and the typicality account are somehow *competing* accounts of entropy-increase and convergence to equilibrium. Witness, for instance, Huw Price who writes with respect to the latter:

> In essence, I think – although he himself does not present it in these terms – what Boltzmann offers is an alternative to his own famous H-Theorem. The H-theorem offers a dynamical argument that the entropy of a non-equilibrium system must increase over time, as a result of collisions between its constituent particles. [...] The statistical approach does away with this dynamical argument altogether. (Price, 2002, p. 27)

Similarly, the pertinent entry in the Stanford Encyclopedia of Philosophy (Uffink, 2017) presents Boltzmann's work as a series of rather incoherent (and ultimately inconclusive) attempts to explain thermodynamic irreversibility.

I am convinced that the reason why Boltzmann did not present the "statistical approach" as an alternative to the H-theorem is that, in fact, it isn't. Understood correctly, there is a clear conceptual continuity between the H-theorem and the typicality account, so that the latter does not appear as a break with Boltzmann's earlier work but as a distillation of its essence (see also Goldstein (2001), Goldstein and Lebowitz (2004)). To make this case, we shall first review the basic setting of the H-theorem and the concept of *distribution functions* and *kinetic equations* more broadly.[11]

We recall that the microstate of an N-particle system is represented by a point $X = (q_1, ..., q_N; p_1, ..., p_N)$ in the $6N$-dimensional phase space Γ, comprising the position and momenta of all particles. The same state (modulo permutations of the particles) can also be represented as N points in the 6-dimensional *μ-space*, whose coordinates correspond to position and velocity of a *single particle*, i.e., $X \rightarrow \{(q_1, v_1), ..., (q_N, v_N)\}$, with $v_i := p_i/m$.

Many results in many-body physics and statistical mechanics, most famously Boltzmanns H-theorem, are concerned with the evolution of a function $f_X(q, v)$ on this μ-space, which provides an efficient description of the most important (macroscopic)

[11] For a good introduction, see also Davies (1977); for more detailed mathematical treatments, e.g., Spohn (1991), Villani (2002).

characteristics of a system in the microstate X. This function is the *empirical distribution* or *coarse-grained density* of points in μ-space. We can think of dividing μ-space into little cells – whose dimension is large enough to contain a great number of particles, yet very small compared to the resolution of macroscopic observations – and counting the number of particles in each cell. For fixed q and v, $f_X(q,v)$ then corresponds to the fraction of particles in the cell around (q,v). In the limit where the size of the cells goes to zero (for fixed N), the coarse-grained empirical distribution becomes the microscopic distribution

$$\mu_X := \frac{1}{N}\sum_{i=1}^{N} \delta(q-q_i)\,\delta(v-m^{-1}p_i). \tag{9.20}$$

It is important to emphasize again that although $f_X(q,v)$ is technically a probability density (just like μ_X is technically a probability measure), there is nothing *random* about it. Instead, we should think of $X \mapsto f_X$ itself as a special type of macro-variable, i.e., a coarse-graining function of microstates. In particular, we may compute the Boltzmann entropy associated to any such distribution. Suppose we divide μ-space into $m \ll N$ cells $(C_1,\dots,C_m)$ and denote by N_k the number of particles in the cell $C_k, k \in \{1,..,m\}$. By simple combinatorics, there are

$$\frac{N!}{N_1!\cdots N_m!} \tag{9.21}$$

ways to distribute the particles over the cells, which lead to the same occupation numbers. And with the Sterling approximation $n! \approx \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, we find for the Boltzmann entropy associated to f_X:

$$S \approx k_B\left(const. - \sum_{k=1}^{m} N_k \log(N_k)\right). \tag{9.22}$$

Writing $N_k = Nf_k|C_k|$ with $|C_k|$ the size (6-dimensional volume) and f_k the density of particles in the cell C_k, we thus have $S \approx const. - Nk_B \sum\limits_{k=1}^{m} |C_k| f_k \log(f_k)$ and see Boltzmann's famous *H-functional*

$$H[f_t] = \int f(t,q,v)\log f(t,q,v) \tag{9.23}$$

emerging in the continuum limit, such that a decrease of $H(t)$ corresponds to an increase of the Boltzmann entropy.

Indeed, the distribution function becomes a truly powerful concept when one considers effective models in which a continuous function $f(t,q,v)$ follows an autonomous time-evolution given by a partial differential equation, a so-called *kinetic equation*, of the form:

$$\partial_t f + p\cdot\nabla_q f + K\cdot\nabla_p f = (\partial_t f)_{\text{coll}}\,. \tag{9.24}$$

Here, K is a force term describing long-range interactions and $(\partial_t f)_{\text{coll}}$ the collision term characteristic of the *Boltzmann equation.* The classical ansatz is

$$(\partial_t f)_{\text{coll}}\,(q,v) = \int W(v_1,v_2;v_3,v)\,[f(t,q,v_1)f(t,q,v_2) - f(t,q,v_3)f(t,q,v)]\,\mathrm{d}v_1\mathrm{d}v_2\mathrm{d}v_3$$

with an appropriate scattering kernel $W(v_1,v_2;v_1',v_2')$, giving the probability per unit time that a collision of two particles with velocities v_1 and v_2 results in velocities v_1' and v_2', respectively.

Important examples of kinetic equations without collision term are *Vlasov*[12] or *mean field* equations with a force term of the form:

$$K(t,q) = -\int \nabla V(q-q')f(t,q',v')\mathrm{d}q'\mathrm{d}v' \tag{9.25}$$

for an interaction potential V. The intuition is thereby that every particle feels the average force exerted by the current particle distribution.

In any case, the continuous distribution $f(t)$ arising as a solution of the kinetic equation (9.24) is supposed to approximate (in the limit of large particle numbers) the actual empirical distribution of the N-particle system evolving according to the pertinent micro-dynamics. To derive the kinetic equation, i.e., justify the effectice model based on the more fundamental microscopic theory, is thus to prove a statement of the following kind: Let $f(t,q,v)$ be a solution of (9.24) with boundary condition $f(0,q,v) = f_0(q,v)$. If the (continuous) density f_0 is a good approximation to the empirical distribution f_X of the initial microstate X, then $f(t)$ will be a good approximation to the empirical distribution $f_{X(t)}$ of the time-evolved microstate $X(t)$.

$$\begin{tikzcd}
f_X \arrow[r, dash, "\approx"] \arrow[d, "\text{microscopic time-evolution}"'] & f_0 \arrow[d, "\text{kinetic equation time-evolution}"] \\
f_{X(t)} \arrow[r, dash, "\approx"] & f(t)
\end{tikzcd} \tag{9.26}$$

For somewhat realistic interactions, this won't be true for all initial configurations X with $f_X \approx f_0$ but only for typical ones (see, e.g., Hauray and Jabin (2015); Lazarovici and Pickl (2017) for pertinent results about mean field equations).

Example (Weak convergence)**.** Mathematically, the relevant approximation is made precise in terms of the weak topology on the space of probability measures. For a

[12] An equation of this type was introduced by A.A. Vlasov in his work on plasma physics Vlasov (1938, 1968) and even earlier by J.H. Jeans in the context of Newtonian stellar dynamics Jeans (1915).

sequence $(\mu_k)_k$ of normalized measures, weak convergence to ν is denoted by $\mu_k \rightharpoonup \nu$ and means that

$$\int \phi(x)\,\mathrm{d}\mu_k(x) \to \int \phi(x)\,\mathrm{d}\nu(x), \quad k \to \infty,$$

for all bounded and continuous functions $\phi : \mathbb{R}^n \to \mathbb{R}$.
A convenient metric inducing this topology is the *bounded Lipschitz distance* defined as:

$$d_{BL}(\mu,\nu) := \sup_{\phi}\Big\{\Big|\int \phi(x)\,\mathrm{d}\mu(x) - \int \phi(x)\,\mathrm{d}\nu(x)\Big| \;:\; \sup_{x\neq y}\frac{\phi(x)-\phi(y)}{|x-y|} = 1,\ \sup_{x}|\phi(x)| = 1\Big\}.$$

Hence, we can understand $f_{X(t)} \approx f(t)$ to mean that the bounded Lipschitz distance between the measures $f_X\,\mathrm{d}^3q\,\mathrm{d}^3v$ and $f(t)\,\mathrm{d}^3q\,\mathrm{d}^3v$ is small, implying approximately equal results when (somewhat well-behaved) "macro-variables" on μ-space are integrated with respect to the empirical distribution f_X and the theoretical distribution $f(t)$, respectively.

However, it is usually convenient for technical reasons to work in the limit $N \to \infty$, and consider a sequence of microscopic systems with increasing particle number whose empirical distributions converge to $f(t)$. Moreover, while f_X depends in the partition of μ-space into cells, the weak topology allows us to compare the continuous density $f(t)$ directly to the discrete microscopic distribution $\mu_{X(t)}$ (cf. eq. (9.20)), making the "step-function" $f_{X(t)}$ dispensable for technical purposes. We nonetheless keep the focus on $f_{X(t)}$, as it makes the coarse-graining nature of the Boltzmannian distribution function more evident.

One can also take another point of view on kinetic equations that deals – in mathematical lingo – with ensembles or "random initial conditions" but is (of course) best understood as aiming at typicality results. This approach considers measures on the N-particle phase space rather than distributions on the reduced μ-space. Suppose that at $t = 0$, the particles are identically and independently distributed according to f_0, that is, in other words, according to the product measure $F_0^N = \otimes^N f_0$ on Γ. If F is evolved with the N-particle flow determined by the microscopic dynamics, one easily checks that it satisfies the *Liouville equation*

$$\partial_t F_t^N + \sum_{i=1}^{N} p_i \cdot \nabla_{q_i} F_t^N + \sum_{i=1}^{N} \frac{1}{N} \sum_{i\neq j} K(q_i - q_j)\cdot \nabla_{p_i} F_t^N = 0. \tag{9.27}$$

Now one would like to establish that under this time-evolution, the particles remain "approximately independent" with $F_t^N \approx \otimes^N f_t$, where f_t is a solution of the kinetic equation (9.24) with initial condition f_0. Formally, this approximation is understood as a weak convergence of marginals. Writing $z_i = (q_i, v_i)$, the reduced k-particle marginal is

$$^{(k)}F_t(z_1, ..., z_k) := \int f_t(Z)\, \mathrm{d}^3 z_{k+1} ... \mathrm{d}^3 z_N, \tag{9.28}$$

and one tries to prove that

$$^{(k)}F_t \rightharpoonup \otimes^k f_t, \quad N \to \infty. \tag{9.29}$$

This is the modern mathematical formulation of *molecular chaos.* It is basically equivalent to the "deterministic" result sketched in Fig. 9.26 for *typical* initial conditions.

Typicality is thereby understood with respect to the product measure $F_0^N = \otimes^N f_0$. At first glance, this might seem to conflict with our insistence that there aren't different competing typicality measures for classical mechanics, but that the Liouville measure, respectively the induced microcanonical measure, is always the appropriate choice. Indeed, in Boltzmannian statistical mechanics, non-equilibrium situations should be first and foremost characterized by a special macrostate. In the present problem, we consider systems starting out in the macro-region $M_{f_0} = \{X \in \Gamma : \mu_X \approx f_0\}$, and for large N, F_0^N is indeed equivalent to the uniform measure restricted to M_{f_0} (in the limit $N \to \infty$, this equivalence is exact in the sense that F_0^N and $\lambda|M_{f_0}$ are absolutely continuous with respect to each other). Because of the manifest statistical independence of the particles, F_0 is just much easier to work with.

If we now recall from (8.11) that the k-particle marginals of the microcanonical measure (for $N \gg k$) are *Maxwellian distributions*, this may already indicate how molecular chaos for the Boltzmann equation could establish convergence to equilibrium: typical initial conditions in M_{f_0} evolve into the equilibrium region $M_{f_{eq}}$ characterized by an (approximately) Maxwellian velocity distribution. We will discuss this in more detail in the next section.

Remark (Scaling limits)**.** The derivation of a kinetic equation always requires an appropriate *rescaling* of the microscopic dynamics to ensure that the relevant physical quantities remain of constant order in the limit $N \to \infty$. Conceptually, this is best understood as a *dimensional* rescaling of the time, position, and/or momentum coordinates. For the Vlasov equation, the relevant regime is the *mean field scaling* $V \to \frac{1}{N}V$, which ensures that the total mass/charge of the system remains of order 1. This corresponds to tracking the time evolution on large (macroscopic) time scales, i.e., in rescaled coordinates $t' = N^{-1/2}t$, $p' = N^{1/2}p$. To derive the Boltzmann equation, one has to make some ansatz for the particle collisions in the microscopic dynamics. The simplest (interesting) one would be the hard spheres model, which, evidently, is itself an idealization. In any case, the relevant scaling regime is the *Boltzmann-Grad* limit in which the scattering radius scales as $r(N) \sim N^{-1/2}$. This is typically realized in rarified gases.

While kinetic equations are commonly and successfully applied in many areas of physics and chemistry, the rigorous justification of the continuous model can be an awfully hard mathematical problem. For the Boltzmann equation, the landmark result of Lanford (1975) establishes molecular chaos only for a very short time interval (a fraction of the particles' mean free time). Subsequent results have extended the proof to a larger class of scattering potentials but not yet overcome this crucial limitation.

Boltzmann's Stoßzahlansatz

Let us now take a more informal look at Boltzmann's equation and H-theorem. The goal of the H-theorem is to show the convergence of an initial non-equilibrium distribution $f_0(q, v)$ to the Maxwell distribution $f_{eq}(q, v)$.

We have already seen from eq. (8.12) that the Maxwell distribution corresponds to the equilibrium state in Boltzmann's sense, i.e., the typical value of the macro-variable $X \to f_X$. In other words, while the coarse-grained distribution f_X will be different for different microscopic configurations X, it is in fact (more or less) the same for the *overwhelming majority* of possible microstates, namely (approximately) of the form

$$f_X(q, v) \propto e^{-\frac{1}{2}m\beta v^2},$$

for a constant β that is the inverse temperature of the system. Note that the distribution having no q-dependence means that the gas is homogeneously distributed over the entire volume, with no correlations between position and velocities, i.e., with *uniform temperature.*

This crucial insight does not appear explicitly in Boltzmann's H-theorem, however, which is rather based on the following three claims:

1) For a low-density gas, the time-evolution of $f_{X(t)}(q, v)$ is well described by an effective equation, now known as the *Boltzmann equation.*

2) For a solution $f(t, q, v)$ of the Boltzmann equation, the *H-function* $H[f(t)] = \int f(t, q, v) \log f(t, q, v) \mathrm{dqdv}$ is monotically decreasing in t. (Whereas $H(f_{X(t)}(q, v))$ for the actual coarse-grained distribution will fluctuate.) Recall that in (9.22), we have already identified the H-function as a (negative) measure of the Boltzmann entropy.

3) The H-function reaches its *minimum* for the Maxwell-distribution $f_{eq}(q, v)$.

   Together with 2), this implies, in particular, that the Maxwell distribution is a *stationary* solution of the Boltzmann equation.

Propositions 2) and 3) are fairly standard mathematical results. The crux of the matter is proposition 1). When Boltzmann first presented the H-theorem in 1872, he argued that a dilute gas *must* evolve in accord with his equation; he later had to mitigate this statement, claiming, in effect, only that it would *typically* do so on empirically relevant time-scales. Indeed, proposition 1), and therefore the H-theorem must be understood as typicality statements.

Boltzmann's original derivation of the Boltzmann equation was famously based on the *Stoßzahlansatz* or the assumption of *molecular chaos*. "Assumption" is, unfortunately, not a perfectly accurate translation of the German word *Ansatz*. Whereas the first is often used synonymously with a logical *premise*, the latter has a distinctly pragmatic character. A better translation would be "working hypothesis" – a plausible (though oversimplified) guess, which is, in the first instance, validated by its success but would ultimately require a deeper justification. So again, we have to keep in mind that Boltzmann's derivation is a brilliant physical argument, but not a rigorous mathematical proof.

In any case, the *Stoßzahlansatz* is an assumption about the *relative frequencies* of collisions between the particles in the gas. Denoting by $\mathcal{N}(t, q; v_1, v_2)$ the number of collisions happening near q in a small time-interval around t between particles with velocity (approximately) v_1 and v_2, the Stoßzahlansatz is:

$$\mathcal{N}(t, q\,; v_1, v_2) \propto N^2 \, f(t, q, v_1) f(t, q, v_2) \, |v_1 - v_2| \; \mathrm{d}t \, \mathrm{d}q \, \mathrm{d}v_1 \mathrm{d}v_2. \tag{9.30}$$

Simply put: The relative frequency of collisions between particles of different velocities occurring in the cell around q is proportional to the density of particles with the respective velocities near the respective position. The scattering probability being proportional to the product of $f(t, q, v_1)$ and $f(t, q, v_2)$ means that particles of different velocities are *statistically independent* as they contribute to the collisions. This is, more specifically, the meaning of *molecular chaos*.

We are not going to repeat Boltzmann's derivation, but it is true as a matter of mathematical fact that *if* and *as long as* the assumption of molecular chaos and hence the *Stoßzahlansatz* are valid, the Boltzmann equation will hold (as a good approximation to the evolution of the empirical distribution under the actual micro-dynamics). The H-theorem thus hinges on the question, if and in what sense the assumption of molecular chaos is justified.

For the purpose of illustration, let's imagine that we could freeze the system at time $t = 0$ and arrange the position and momentum of every single particle before letting the clock run and the system evolve in time. (Note that there is no issue here as to whether we let the clock run "forwards" or "backward" – the problem is symmetric with respect to the time-evolution in both directions.) Which particles are going to collide and how they are going to collide is then completely determined by these initial conditions and the microscopic laws of motion. We could, for instance, arrange the initial configuration in such a way that "slow" particles will almost exclusively scatter with other "slow" particles, and "fast" particles with other "fast" particles. But such initial conditions are obviously very special ones. For *typical* microscopic configurations, coarse-graining to the initial distribution $f_0(q, v)$, we will find that the relative frequencies with which particles of different velocities meet for the first collision is roughly proportional to the density of particles with the respective velocities, i.e., given by eq. (9.30). This is nothing more and nothing less than the *law of large numbers*. The validity of (9.30) *at the initial time* is thus, like all LNN results, a *typicality statement* and, as such, another mathematical fact.

The critical issue is whether molecular chaos *propagates* with the microscopic dynamics. Assume that after an (infinitesimal) time-interval Δt, for which the Boltzmann-equation is valid, the continuous distribution has evolved into $f(\Delta t, q, v)$. How do we know that (9.30) is still a good approximation for all but a small set of initial conditions? It is still true that eq. (9.30) is satisfied for typical microscopic configurations realizing the *current* distribution, i.e., counting all possible configurations that coarse-grain to $f(\Delta t, q, v)$. But we cannot count all these configurations since the relevant microstates must have evolved from the macro-region realizing the initial distribution $f_0(q, v)$. Mathematically, this constraint translates into a loss of statistical independence at times $t > 0$, making it prima facie questionable whether a law-or-large-numbers statement for the collisions, i.e., (9.30), still holds. This is, by the way, the only meaningful sense in which interactions *build up correlations* (notably, in both time-directions), and it should be distinguished from naive causal intuitions that two particles are somehow intrinsically independent before – but not after – they collide.

Boltzmann's *Stoßzahlansatz* is thus the assumption that statistical independence is sufficiently well preserved under the microscopic time-evolution, or, in other words, that the relative frequency of collisions is always the typical one with respect to the current distribution function.

We have already noted that a rigorous derivation of the Boltzmann equation would amount to a proof of this assumption, i.e., a proof that molecular chaos propagates (in the Boltzmann-Grad limit). This would be a monumental mathematical achievement, a sure claim to fame and a Fields Medal (for anyone young enough to qualify). However, based on physical intuition and various encouraging results, there is little doubt that Boltzmann's assumption – though idealized – is justified. Given that the microscopic dynamics are very chaotic, that the number of particles in a gas is huge, and the gas (by assumption) very dilute so that problematic re-collisions are extremely rare, it is highly plausible that the relative frequencies of scatterings should not become too special – in the sense of deviating significantly from the expectation values (9.30) – unless the initial micro-configuration itself were very special.

Of course, it is important to note that, unless one considers the thermodynamic limit of infinitely many particles, molecular chaos and equation (9.30) will hold at best *approximately* for *all but a small set* of "bad" initial conditions; that this approximation will get worse with time, and that the approximation is only good enough until it isn't. Eventually, a typical system will exhibit sizable fluctuations out of equilibrium, at which point its evolution is no longer adequately described by the Boltzmann equation.

Example (A simple toy-model for the Boltzmann equation)**.** We consider a system of $N \gg 1$ balls. At the beginning, n of the balls are black and $m = N - n$ are white.

When two like-colored balls collide, they change their color; otherwise they stay the same:

$$\begin{aligned} w + w &\to b + b \\ b + b &\to w + w \\ b + w &\to b + w \end{aligned} \tag{9.31}$$

Note that these dynamics are time-symmetric. For our model, we consider discrete time-steps, assuming that in each round a total of $k \ll N$ collisions occur. Now we make the following "Stoßzahlansatz": The probability that a black/white ball enters a collision corresponds to current the fraction of black/white balls in the system. The expectated numbers of collisions in each round are thus:

$$\begin{aligned} &k \cdot \left(\frac{n}{N}\right)^2 \text{ collisions } b + b \\ &k \cdot \left(\frac{m}{N}\right)^2 \text{ collisions } w + w \\ &k \cdot 2\left(\frac{n}{N}\right)\left(\frac{m}{N}\right) \text{ collisions } b + w \text{ resp. } w + b. \end{aligned}$$

Consequently, the expected change in the total number of black and white balls is

$$\begin{aligned} n &\to n + k \cdot 2\left[\left(\frac{m}{N}\right)^2 - \left(\frac{n}{N}\right)^2\right] \\ m &\to m + k \cdot 2\left[\left(\frac{n}{N}\right)^2 - \left(\frac{m}{N}\right)^2\right], \end{aligned}$$

and taking the difference:

$$(n - m) \to (n - m) - \frac{4k}{N^2}(n - m)(n + m) = \left(1 - \frac{4k}{N}\right)(n - m).$$

This is iterated in each round. For large N, typical evolutions will be close to this theoretical expectation and hence, after $T \in \mathbb{N}$ time-steps:

$$(n - m)(T) \approx \left(1 - \frac{4k}{N}\right)^T (n - m)(0) \to 0,\ T \to \infty. \tag{9.32}$$

We thus have *convergence to equilibrium*: the number of black and white balls in the system tends towards the equidistribution; and if we start with an unequal number of black and white balls – i.e., in non-equilibrium – the time-symmetric scattering dynamics (9.31) lead to the irreversible macro-evolution (9.32). However, small deviations from the expectation values will add up over time, leading to fluctations out of equilibrium. At that point, the effective equation (9.32) is no longer valid.

H-theorem as a typicality statement

With all that said, let us summarize once again why Boltzmann's H-theorem is not an alternative way of explaining thermodynamic behavior but a concrete exemplification of the general typicality account.

While the micro/macro distinction does not appear as prominently in the formulation of the H-theorem, it is essential that the function $f(t, q, v)$ pertains to a *coarse-grained* description of the system, hence distinguishing a macro-region in phase space consisting of all microscopic configurations whose distribution is well-approximated by the same f. Convergence to equilibrium is then established for *typical initial conditions* relative to this initial non-equilibrium region. And the equilibrium state – characterized by the Maxwell distribution to which a non-equilibrium distribution typically converges – is, as always, distinguished by the fact that it is the one realized by an *overwhelming majority* of all possible microstates.

Despite the common focus on the *Stoßzahlansatz*, I submit that the tendency to equilibrium is mostly explained by this dominance of the equilibrium state. The explanatory role of molecular chaos is somewhat subsidiary to this insight, namely to validate the intuitively obvious fact that the "most likely" evolutions will thus carry a non-equilibrium configuration into the overwhelmingly large equilibrium region.

Finally, we understand that the *irreversibility* of the Boltzmann equation (as an effective description of a system's macro-evolution) is – as usual – a consequence of the fact that non-equilibrium configurations converging to equilibrium are *typical* with respect to the corresponding macrostate, whereas microstates leading to the time-reversed evolution are *atypical* relative to the equilibrium state, i.e., relative to all micro-configurations coarse-graining to $f_{eq}(q, v)$. The same holds true with respect to any macrostate along the way.

It is often claimed and criticized that the Stoßzahlansatz is a manifestly time-asymmetric assumption, in that the *incoming* rather than the *outgoing* velocities are assumed to be independently distributed according to the current density function (see, e.g., Uffink (2007, p. 117)). This claim, though technically correct, is missing the point, and the misunderstanding seems to be mostly due to the failure to recognize molecular chaos, respectively the Stoßzahlansatz, as typicality statements.

For typical initial conditions (relative to the current macrostate), eq. (9.30) is equally valid for the time-evolution in *both time-directions*. The origin of the asymmetry is, as always, the special boundary condition, i.e., the assumption of a non-equilibrium initial distribution. Relative to this low-entropy initial state, the relevant micro-configurations are typical ones for which entropy increases (in both time-directions), whereas the higher-entropy configurations along the corresponding solutions are necessarily atypical, relative to their current macrostate, with respect to their evolution towards the low-entropy boundary condition. In particular, this atypical evolution towards the past (for which molecular chaos doesn't hold) is explained by the non-equilibrium boundary condition.

Molecular chaos thus breaks time-symmetry only in the good and necessary sense that it applies to the thermodynamic evolution but not to the reversed motion. This does not mean, however, that any time-asymmetry was smuggled into Boltzmann's argument in addition to the one introduced by the non-equilibrium initial state. If the terms "incoming" and "outgoing" velocities are misleading here, we can simply replace them with velocities "towards," respectively "away from" (in a temporal sense) the non-equilibrium boundary condition. Since statistical independence is a typicality property, molecular chaos can only hold for the time-evolution away from the low-entropy boundary condition. For evolutions towards it, the boundary constraints will naturally impose strong, seemingly conspiratorial correlations.

The deeper question, why the low-entropy boundary conditions that we are able to prepare are always "past" rather than "future" ones (and thus why only the Boltzmann rather than the "anti-Boltzmann" equation is actually relevant) goes beyond the scope of the H-theorem per se. It is a question about the *arrow of time* and the boundary conditions of our universe that we will address in Chapter 11.

9.8 Boltzmann vs. Gibbs

Throughout our discussion, we have followed the Boltzmannian approach to statistical mechanics and did not say much about the other influential framework that goes back to J.W. Gibbs. The key difference is often characterized as one between an *individualist* and an *ensemblist* view. In Boltzmannian statistical mechanics, we assign microstates and macrostates to individual systems. This sometimes raises the question, how probability theory can be applied (without resorting to epistemic probabilities), but the question is answered by the concept of typicality. In Gibbsian statistical mechanics, probability measures are interpreted as ensemble distributions and thus taken to describe the state of an (usually hypothetical) ensemble. This often raises the question, what Gibbsian predictions imply for observations on individual systems.

In recent years, the relationship between the Boltzmannian and Gibbsian framework has been a subject of great interest to philosophers of physics (foor a good recent discussion – by a mathematical physicist – see Goldstein (2019)). In this section, I am primarily going to address one contribution that I find less helpful, while trying to make some clarifying remarks of broader relevance along the way.

Charlotte Werndl and Roman Frigg (2017) address, in particular, the question, if and when Boltzmann and Gibbs yield equivalent predictions for equilibrium values of macroscopic observables. In the Boltzmannian framework, the macro-variables take (approximately) constant values on the equilibrium region of phase space, which are thus revealed by a suitable measurement on a system in equilibrium (a system, that is, whose actual microstate is in the equilibrium state). In the Gibbsian framework, equilibrium is a property of an *ensemble*, represented by a stationary distribution ρ on phase space Γ, and it is often (though maybe somewhat carelessly) said that the prediction for a measurement of a macro-variable f on an individual ensemble system is given by the *phase average*

$$\langle f \rangle = \int_\Gamma f(z)\rho(x)\,\mathrm{d}x\,, \tag{9.33}$$

where $x \in \Gamma$ are the phase space coordinates. This quantity is also called the *ensemble average* or *expectation value* of f.

There are many situations in which the Gibbsian phase average agrees – within appropriate error bounds – with the Boltzmannian equilibrium value. Werndl and Frigg mention a criterion which they call the "Khinchin condition" and which they characterize briefly as the phase function having "small dispersion for systems with a large number of constituents." Indeed, in less technical terms, a sufficiently small dispersion of the macro-variable means precisely that typical values (the Boltzmann equilibrium value) are close to the average value (the Gibbsian equilibrium value). Another way to formulate the Khninchin condition – now from a Boltzmannian perspective – is to say that there exists a unique Boltzmann equilibrium whose corresponding macro-region exhausts almost the entire phase space volume in terms of the pertinent stationary measure. Formally:

$$\mu_\rho\left(\Gamma_{\mathrm{eq}}\right) = \int_\Gamma \mathbb{1}\{f(x) \in (\xi \pm \Delta\xi)\}\,\rho(x)\,\mathrm{d}x = 1 - \epsilon, \tag{9.34}$$

where $\Delta\xi$ is very small compared to ξ, and ϵ is very small compared to 1 (and $\mathbb{1}_A$ denotes the characteristic function of the set A). For then, the macro-variable f takes an (approximately) constant value – the Boltzmannian equilibrium value $\xi \pm \Delta\xi$ – on a set of measure close to 1 – the Boltzmannian equilibrium region Γ_{eq}. Hence, the phase average (9.33) will be close to the Boltzmannian equilibrium value (provided f is somewhat well-behaved, and its values don't suddenly "explode" outside the equilibrium region). Rigorously:

$$\begin{aligned} |\langle f \rangle - (1-\epsilon)\xi| &\leq \int_{\Gamma_{\mathrm{eq}}} |f(x) - \xi|\,\rho(x)\,\mathrm{d}x + \int_{\Gamma\setminus\Gamma_{\mathrm{eq}}} |f(x)|\rho(x)\,\mathrm{d}x \\ &\leq (1-\epsilon)\Delta\xi + \epsilon \sup_{x\in\Gamma\setminus\Gamma_{\mathrm{eq}}} |f(x)|, \end{aligned} \tag{9.35}$$

and thus $\langle f \rangle \approx \xi$ assuming $\epsilon \sup_{x\in\Gamma\setminus\Gamma_{\mathrm{eq}}} |f(x)| \ll |\xi|$.

It is important to emphasize that, unless one considers a thermodynamic limit, "the Boltzmannian equilibrium value" refers, in general, to a small range of values of f. This kind of coarse-graining is both essential for probabilistic estimates and physically called for, considering that the relevant measurements have limited accuracy. For Werndl and Frigg, in contrast, every single value of the macro-variable f defines a different "macrostate," implying that, for them, even the slightest variation in the respective physical quantity – pressure, density, energy, etc. – leads to a system being "out of equilibrium." What the authors call the "Boltzmann equilibrium" is thus not the equilibrium as defined by Boltzmann or used in Boltzmannian statistical mechanics, and it is unfortunate, since misleading, that they refer to it by the same name. For our further discussion, we will thus refer to it as the "Werndl-Frigg equilibrium" instead.

The existence of a dominant Boltzmann equilibrium (in the sense of Boltzmann) is, in fact, the generic case in statistical mechanics, implying that Boltzmann and Gibbs make (in general) equivalent predictions for systems in the respective equilibria. If we agree that the Boltzmannian formulation is the more fundamental one, this also explains *why* Gibbsian phase averaging yields relevant predictions for individual measurements. Simply put, the macro-variables are essentially constant across most of the ensemble and the average reflects this typical value.

It is a common misconception – repeated in many textbooks – that the empirical relevance of the phase average is due to *Birkhoff's ergodic theorem*, which establishes equality between (9.33) and the *time-average* $\lim\limits_{T\to\infty} \frac{1}{T}\int_0^T f(x(t))\mathrm{dt}$ for almost all initial conditions. The argument is that one measures ergodic time-averages because measurements are not instantaneous but require a prolonged interaction between apparatus and system. This is completely wrong since ergodic time scales are *much* too long (see Goldstein (2001)).[13]

On the other hand, there are famous and well-studied cases in which the Khinchin condition (9.34) doesn't hold. For instance, in the two-dimensional Ising model without external field, it makes sense to speak of *two* Boltzmann equilibria below the critical temperature, corresponding to a positive or negative magnetization, respectively. The distribution ρ is, however, symmetric under a flip of all spins, hence yielding an average magnetization of zero. There is nothing inconsistent or mysterious about this fact, as long as we keep in mind that the Gibbsian value refers, in the first place, to an ensemble average. In particular, in statistical mechanics, one does not try to draw interesting conclusions about the Ising model from such phase averages. Instead, one usually studies so-called phase transitions at the critical temperature by fixing either $+1$ or -1 boundary conditions (referring to the polarization of spins at the edge of the lattice), thus implicitly picking one of the two magnetization states.

Boltzmann vs. Werndl and Frigg

Werndl and Frigg (2017) mention the magnetization in the Ising model only briefly but present instead other examples for which they claim a disagreement between Boltzmannian and Gibbsian equilibrium values. One such example is based on the "baker's gas," a mathematical model for the ideal gas that the authors have used in various publications to argue that the Boltzmann equilibrium (the largest macro-region) fails to be *dominant*, i.e., does not exhaust a majority of the phase space volume as stated in equation (9.34). To this end, the authors partition the one-particle phase space

[13] Schwarz (1992, p. 23-24) put it most succinctly: "*[T]he delicious ingenuity of the Birkhoff ergodic theorem has created the general impression that it must play a central role in the foundations of statistical mechanics. [...] The Birkhoff theorem in fact does us the service of establishing its own inability to be more than a questionably relevant superstructure upon [the] hypothesis [of absolute continuity].*"

(corresponding to the unit square in the baker's model) into k cells and claim that the macrostate corresponding to the Boltzmann equilibrium is the "uniform distribution" for which each cell contains exactly $\frac{N}{k}$ particles. It is then easy to see that while the phase space volume associated with this distribution is greater than the phase space volume associated with any other particular arrangement of particles over the cells, it will not exhaust a majority of phase space volume for large N. In their paper, the authors exploit this fact – amounting to an apparent violation of the Khinchin condition (9.34) – by introducing an artificial macro-variable, weighing the previously defined "macro-regions" in such a way that the Gibbsian phase average differs significantly from the value associated with the uniform distribution.

However, as emphasized above, the authors' reference to the "Boltzmann equilibrium" is a misnomer, since they use a notion of equilibrium that does not correspond to the concept introduced by Boltzmann and used in Boltzmannian statistical mechanics (thereby repeating the arguments of Lavis (2005) that we have already criticized above). *Exact* uniform distributions, where each cell contains exactly $\frac{N}{k}$ particles, are very special configurations, their measure actually goes to zero for large N. But configurations for which the fraction of particles contained in each cells differs only slightly from $\frac{1}{k}$ are macroscopically indistinguishable and coarse-grain to the same macrostate in Boltzmann's sense. Otherwise, we would have to say, for instance, that a gas is "out of equilibrium" if the left-hand-side of the volume contains even a single particle more than the right-hand-side.

Compare this with Boltzmann's discussions of the Maxwellian velocity distribution as the equilibrium state of an ideal gas. Here, Boltzmann was very explicit about the fact that "for a finite number of molecules, the Maxwell distribution will not hold exactly but only to a good approximation." (Boltzmann, 1896b, translation D.L.)

In the case of the baker's model, the dominant Boltzmann equilibrium contains all configurations for which the relative number of particles in each cell is within $\frac{1}{k} \pm \frac{const.}{\sqrt{N}}$. Since $\frac{1}{\sqrt{N}}$ is a tiny number for macroscopic N, corresponding to density fluctuations of less than one-tenth of a billionth of a percent, these configurations look *macroscopically* uniform and constitute the relevant equilibrium state. The Khinchin condition (9.34) is thus satisfied, and the Boltzmannian and Gibbsian equilibrium values will be equivalent for all sensible macro-variables.

Law of Large Numbers vs. EET

By the same token, all other examples presented by Werndl and Frigg may show a disagreement between the Gibbsian equilibrium and their own, but shed no light on the relationship with the real Boltzmann equilibrium. Nevertheless, the authors go on to conclude that it is "[a]n important task of the foundations of SM [statistical mechanics] ... to classify under which conditions the two frameworks lead to the same results and under which conditions they do not" (p. 1300) and present a "new theorem specifying a set of conditions" under which "Boltzmannian" and Gibbsian equilibrium values coincide. I quote their result in full:

> **Equilibrium Equivalence Theorem (EET):** Suppose that the system (X, T_t, μ_X) is composed of $N \geq 1$ constituents. That is, the state $x \in X$ is given by the N coordinates $x = (x_1, ..., x_N)$; $X = X_1 \times X_2 \ldots \times X_N$, where $X_i = X_{oc}$ for all $1 \leq i \leq N$ (X_{oc} is the one-constituent space). Let μ_X be the product measure $\mu_{X_1} \times \mu_{X_2} \ldots \times \mu_{X_N}$, where $\mu_{X_i} = \mu_{X_{oc}}$ is the measure on X_{oc}. Suppose that an observable κ is defined on the one-particle space X_{oc} and takes the values $\kappa_1, \ldots, \kappa_k$ with equal probability $1/k, k \leq N$. Suppose that the macro-variable K is the sum of the one-component observable, i.e., $K(x) = \sum_{i=1}^{N} \kappa(x_i)$. Then the value corresponding to the largest macro-region as well as the value obtained by phase space averaging is $\frac{N}{k}(\kappa_1 + \kappa_2 + \ldots \kappa_k)$.

This theorem tries to identify the Werndl-Frigg equilibrium value with the Gibbs average. It does not relate to the Boltzmann equilibrium. That is because the "largest macro-region" of Werndl and Frigg refers to the set on which the macro-variable takes *exactly* the average value, and this is, again, in contrast to the Boltzmannian framework, in which a small range of (for all practical purposes indistinguishable) values coarse-grains to one and the same macrostate.

It should be noted, however, that part of the conditions specified by Werndl and Frigg are sufficient (though by no means necessary) for the equivalence of Gibbsian and Boltzmannian equilibrium values. In fact, under these conditions, the equivalence is a standard exercise in statistical mechanics, based on the *law of large numbers* (LLN) – the fundamental theorem of probability theory that we have already discussed in Chapter 2. The LLN yields for a family of independent and identically distributed random variables:

$$\mu\left(\left\{x : \left|\frac{1}{N}\sum_{i=1}^{N}\kappa(x_i) - \frac{1}{k}(\kappa_1 + \kappa_2 + \ldots + \kappa_k)\right| < \epsilon\right\}\right) \geq 1 - \frac{\sigma^2}{\epsilon^2 N}\,, \tag{9.36}$$

for any $\epsilon > 0$, where σ^2 is the variance of κ. For the macro-variable $K(x) = \sum_{i=1}^{N} \kappa(x_i)$ then, which is extensive and growing with N, we can set $\epsilon = N^{-\delta}$ for $\delta \in [0, \frac{1}{2}]$ so that, in terms of K (and writing $\overline{K} := \frac{N}{k}(\kappa_1 + \kappa_2 + \ldots + \kappa_k)$ as a "phase average"), the LLN estimate becomes

$$\mu\left(\left\{x : \left|K(x) - \int K(x')\,\mathrm{d}\mu(x')\right| < N^{1-\delta}\right\}\right) \geq 1 - \frac{\sigma^2}{N^{1-2\delta}}\,. \tag{9.37}$$

Note that the bound $N^{1-\delta}$ is small compared to $\overline{K}$, which is of order N, i.e., it is the *relative* deviation $\left|\frac{K-\overline{K}}{\overline{K}}\right| \lesssim \frac{N^{1-\delta}}{N} = N^{-\delta}$ that becomes vanishingly small for large particle numbers. In particular, for a macroscopic system, we have $N \sim 10^{24}$ (from Avogadro's constant), and setting $\delta = \frac{1}{3}$, we can conclude that K deviates from $\overline{K}$ by less than one millionth of a percent on a set of measure (approximately) $0,999999$.[14]

To sum up in less technical terms, the weak law of large numbers says precisely that for large N (which is the relevant case in statistical mechanics), phase space is dominated by an equilibrium region, on which the value of the macro-variable K is very close to the expectation value (= phase average). And this is precisely the empirical equivalence of Boltzmannian and Gibbsian equilibrium values. The LLN also yields immediately the Khinchin condition, both in the sense of small dispersion (which is actually how the LLN is usually proven) and in the form of equation (9.34) (to be compared with (9.37)). Finally, the LLN holds even under much more general assumptions than those of Werndl's and Frigg's EET, namely (in the usual textbook version) for any sum of uncorrelated and identically distributed random variables.

I have to emphasize again that while the LLN applies immediately under their stated assumptions, the theorem proven by Werndl and Frigg is not a LLN statement because the authors have a different notion of "Boltzmann equilibrium" in mind. The LLN, like much of statistical mechanics, is about estimates. Werndl and Frigg don't do estimates. Instead, in their paper, the Werndl-Frigg equilibrium value and its equality to the phase average is supposed to be *exact*. This is why, in addition to considering standard conditions for the LLN, Werndl and Frigg assume a particularly simple distribution of the macro-variable for which the average coincides with the most likely value. The EET is then a straightforward exercise in combinatorics. Its physical relevance, however, is questionable, to say the least. First, the measure of what the authors call the "largest macro-region," i.e., the set on which K takes *precisely* the value $\frac{N}{k}(\kappa_1+\kappa_2+\ldots+\kappa_k)$, actually goes to *zero* for large N. Interpreting this measure probabilistically, it is thus extremely unlikely for a system to be in this Werndl-Frigg equilibrium. Second, while the macro-variables assumed in the EET are discrete, different values can be very close for large N. Therefore, many different "macrostates" in the sense of Werndl and Frigg will be empirically indistinguishable, given the limited resolution of measurements on macroscopic systems. This is a crucial difference between macrostates in the sense of Werndl and Frigg and macrostates in the sense of Boltzmann. In general, it should be clear that the relevant notion of equivalence for Boltzmannian and Gibbsian equilibrium predictions can only be *empirical equivalence*, i.e., that the respective values agree to a sufficiently good approximation.

The law of large numbers is, in fact, the paradigm that we should have in mind when we think about the Boltzmann equilibrium: there is a certain *range* of typical values for the relevant quantities; and the larger the particle number N, the more weight (phase space measure or probability, if you wish) is concentrated on an ever smaller range of values around the mean.

It is also a standard result in probability theory that the variance (= dispersion squared) for a sum of independent random variables (as considered by Werndl and

[14] A tacit assumption, generally made, is that the variance σ^2 of the *one-constituent variables* κ_i is of order 1. If σ is extremely large, or somehow chosen to increase with N, the LLN may fail to provide relevant estimates, though such cases seem unphysical.

Frigg) is additive. This is to say, in particular, that typical fluctuations are of the order $\sqrt{N}$ and we will not have a dominant equilibrium region if the coarse-graining into macrostates is finer than that (cf. also the *central limit theorem* discussed in Ch. 2.2). This is a simple mathematical fact, not a foundational problem.[15]

To be clear: a macro-variable, qua mathematical object, is usually some nice function of the microscopic variables – think for instance of the energy $H(q,p)$ as a function of the particles' positions and momenta in a canonical ensemble. But such a variable is, in general, too fine-grained to consider all its different values as macroscopically distinct. There is thus not a one-to-one correspondence between the possible values of the macro-variable and Boltzmannian macrostates. Boltzmannian macrostates are supposed to be observationally distinguishable by relevant means.

While the Boltzmannian equilibrium value is thus necessarily "unsharp," the Gibbsian equilibrium value, if identified with the phase average (9.33), is a definite real number by definition. It would be a mistake, however, to read this as an "infinitely precise" prediction of the Gibbsian theory, as if it said that the observed value in every single instance is exactly the mean. Instead, one can, for instance, compute the ensemble variance

$$(\Delta f)^2 := \int (f(x) - \langle f \rangle)^2 \rho(x)\, \mathrm{d}x \tag{9.38}$$

and identify the Gibbsian prediction (in the sense of a typicality statement), to be compared with the Boltzmannian one, with $\langle f \rangle \pm \Delta f$.

As to the broader question under which conditions Boltzmannian and Gibbsian equilibrium predictions are equivalent, a good case can be made that the Khinchin condition, in the sense of "uniqueness and dominance of the Boltzmann equilibrium," is not only sufficient[16] but also necessary. For if the condition is violated, we have either *no* Boltzmann equilibrium – and thus no Boltzmann equilibrium value – or multiple Boltzmann equilibria, so that the Gibbsian phase average will correspond to an average of the Boltzmann equilibrium values rather than any one in particular (unless, of course, this average happens to be itself among the set of equilibrium values). However, instead of arguing this point in greater detail, a more relevant observation is the following: If the Khinchin condition doesn't hold, it means that there's a high probability of finding macro-values that differ significantly from the Gibbsian phase average, so that this phase average, as a prediction for individual measurements, is highly dubious in the first place.

Mind the gap

Werndl and Frigg have written a series of papers (e.g., 2015a,b, 2017) attacking the premise of a dominant Boltzmann equilibrium and addressing the problems – like the

[15] Note that partitioning the one-constituent space into cells is only a coarse-graining on the microscopic scale (order 1 for an extensive variable) and thus not sufficient for a macroscopic coarse-graining.

[16] Together with an appropriate bound on the variation of the macro-variable as specified after eq. (9.35).

non-equivalence of Boltzmannian and Gibbsian equilibrium values – that ensue. In almost every case, these alleged foundational problems arise only from an idiosyncratic definition of "macrostates" and "Boltzmann equilibrium" that does not correspond to the relevant Boltzmannian concepts. If one fails to consider an appropriate coarse-graining of the microscopic state space – as suggested both by physical considerations and elementary results in probability theory – the resulting "counterexamples" to the existence of a dominant "equilibrium state" are neither surprising nor relevant to Boltzmann's statistical mechanics. No Boltzmannian – least of all Boltzmann himself – ever claimed that one could partition phase space in any arbitrary matter and end up with a dominant equilibrium state.

In general, I am skeptical of this *ad absurdum* approach to the foundations of statistical mechanics – constructing artificial counterexamples that create artificial problems. In my view, it misses the point of the discipline, which is not an axiomatic theory but an effective framework for the description of complex systems which requires some degree of pragmatism and good physical sense. A crucial difference is that in an axiomatic theory, any counterexample can point to foundational issues, while in statistical mechanics, some counterexamples point merely to inadequate use.

Chapter 10

Causality and the Arrow of Time

> The law of causality, I believe, like much that passes muster among philosophers, is a relic of a bygone age, survivinig, like the monarchy, only because it is erroneously supposed to do no harm.
>
> — Betrand Russel, On the notion of cause, 1912.

A standard example of a causal relation is: "The ball hitting the window is the cause of the window breaking." However, it is not true that a ball (with this and that momentum) hitting the window (of this and that constitution) will *necessarily* break the window. There are certainly microscopic initial conditions realizing the respective macrostate for which the glass will resist the impact. What is very plausibly true is that *typical* micro-configuration, coarse-graining to a window and a ball flying towards it, will evolve into configurations coarse-graining to a broken window and a ball on the other side.

With this in mind, I propose the following typicality analysis of causation. We shall say that a macrostate (macro-event) A *causes* another macrostate B under the conditions C if

$$\neg\texttt{Typ}(B \mid C) \text{ but } \texttt{Typ}(B \mid A \cap C). \tag{10.1}$$

Notably, this is a relation between macrostates of a (closed) physical system characterized by the properties A and B and the "ceteris paribus" clause C. It is not a relation between ball and window conceived as separate physical systems since we must represent both on the same phase space.

Our definition incorporates a sort of "Hume counterfactual," which is not *if A had not happened, B would not have happened* – the truth value of this counterfactual is actually underdetermined if A and B are characterized in macroscopic terms – but: *if A had not happened, B happening would not have been typical.* In many interesting cases, even $\texttt{Typ}(\neg B \mid C)$ and thus the stronger counterfactual *if A had not happened, B would typically not have happened* are true, but I don't consider this to be necessary. One may very well cause events that would not have been otherwise atypical, e.g., by cheating to make sure that one decisive die roll lands on *six.* (Whereas, according to the proposed analysis, one cannot cause events that will typically occur anyway.)

We note that the following inferences are not generally true:

$$\texttt{Typ}(E|A), \texttt{Typ}(E|B) \Rightarrow \texttt{Typ}(E|A \cap B) \tag{10.2}$$

$$\texttt{Typ}(E|A), \texttt{Typ}(E|B) \Rightarrow \texttt{Typ}(E|A \cup B) \tag{10.3}$$

Counterexamples are schematically sketched in Fig. 10.1.

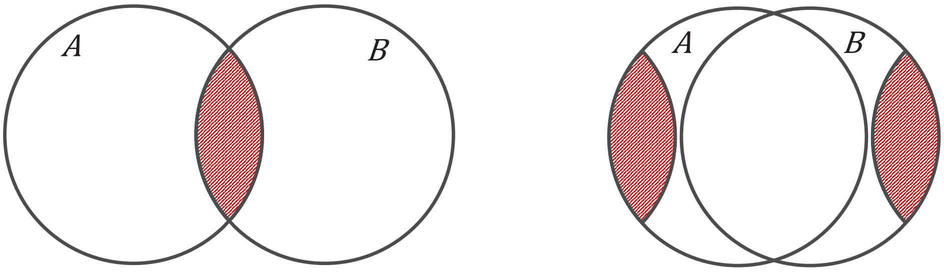

Figure 10.1: Schematically: "bad" configurations, for which E fails to obtain, are red. Left: the red region is small in A and in B but not in the intersection. Right: the red region is small in A and in B but not in the union (it makes up a larger proportion of the area of $A \cup B$).

At first, I found the failure of (10.3), in particular, counterintuitive and worried that the typicality definition of causation thus couldn't be correct. If A causes E and B causes E, shouldn't this imply that A OR B causes E since at least one of the causes occurs? Further reflection convinced me otherwise. Note that $A \cup B$ is equivalent to $A \cup (B \setminus A)$. But the instances of B that lead to E may also be instances that typically realize A, as well. For example, the propositions "An uncontrolled plane crash causes death" and "Severe physical injury causes death" may both evaluate as true, but *severe physical injury* OR *an uncontrolled plane crash without severe injury* does not cause death.

Similarly (to stay with the morbid examples) it may be true that an overdose of blood pressure-lowering medication causes death and that an overdose of blood pressure-increasing medication causes death but that an overdose of both at once does not. Hence, it seems right that (10.2) can fail, as well.

Inference (10.3) does hold if $E \cap A \cap B = \emptyset$, that is, intuitively, if A and B bring about E in distinct ways. Since then, we have for any measure μ:

$$\mu(E \mid A \cup B) = \frac{\mu(E \cap (A \cup B))}{\mu(A \cup B)} = \frac{\mu(E \cap A) + \mu(E \cap B)}{\mu(A \cup B)} \geq \frac{\mu(E \cap A) + \mu(E \cap B)}{\mu(A) + \mu(B)},$$

so that $\mu(E \mid A) = \frac{\mu(E \cap A)}{\mu(A)} > 1 - \epsilon$ and $\mu(E \mid B) = \frac{\mu(E \cap B)}{\mu(B)} > 1 - \epsilon$ implies

$$\mu(E \mid A \cup B) = \frac{\mu(E \cap A) + \mu(E \cap B)}{\mu(A) + \mu(B)} \geq \frac{(1 - \epsilon)(\mu(A) + \mu(B))}{\mu(A) + \mu(B)} = 1 - \epsilon.$$

Hence, the measure of "bad" configurations, which don't bring about E, doesn't increase under disjunction of disjoint causes.

10.1 Causal Explanations as Typicality Explanations

Hence, upon the view proposed here, causal relations are not fundamental and not instantiated by microscopic interactions per se. Dynamical laws yield a relation of entailment (or maybe necessitation) between physical states at different times, and this relation is symmetric if the laws are bi-deterministic. Given the complete dynamical state of the world at any time t (or, relativistically, on a *Cauchy hypersurface*) the laws determine the complete state of the world at any other time, earlier or later.

Causal relations, on the other hand, are understood in terms of typical macro-evolutions between macrostates. And this relations will manifest asymmetrically in systems which (like our universe) have a thermodynamic arrow of time. Indeed, it follows from the Boltzmannian analysis that in a system with a thermodynamic arrow, the evolution towards the future (the direction of entropy increase) looks like a *typical* one relative to any intermediate macrostate, while the actual microstate is necessarily atypical with respect to its evolution towards the entropic past. This is essentially the reversal of the familiar "paradox" that entropy increase in *both* time directions comes out as typical relative to any non-equilibrium macrostate (see Fig. 10.2).

Thus, in a universe with a thermodynamic arrow, causal relations in the sense of (10.1) will in general only be instantiated between *past causes* and *future events* (with respect to the direction of entropy increase) and causal inferences of the form

$$A, \, \mathtt{Typ}(B \mid A) \rightsquigarrow B \tag{10.4}$$

will only be successful for *predictions*, i.e., when B lies in the entropic future of A.

Of course, in the usual way of speaking, it is possible to cause a lower-entropy state, e.g., when a freezer causes water to freeze into an ice cube. To apply our analysis, however, we have to look at the bigger picture: A room containing a freezer with water typically evolves into a room with a freezer, an ice cube, and somewhat increased temperature. The room may be considered as a closed system for the purpose of this analysis, but the system refrigerator + water alone may not. The emission of heat into the environment must be taken into account to see the thermodynamic arrow.

Since in a universe with a thermodynamic arrow only the macro-evolution into the future is typical relative to the present macrostate, a sensible way of making *retrodictions*, i.e., inferences about the past, is not (10.4) (from a present "cause" to

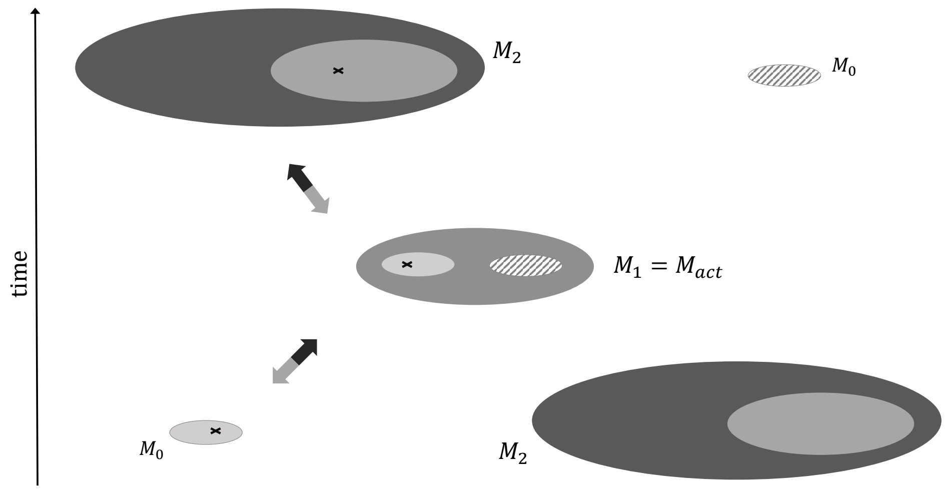

Figure 10.2: We assume, for simplicity, that all macrostates are invariant under the time-reversal transformation ($(q,p) \to (q,-p)$ in classical mechanics). Typical microstates in the intermediate macro-region $M_1 = M_{act}$ evolve into a higher-entropy region M_2 in both time directions. Only a small subset of microstates (light grey area) have evolved from the lower-entropy state M_0 in the past; an equally small subset (shaded area) will evolve into M_0 in the future. The actual microstate (cross) has evolved from the lower-entropy state in the past; only its future time-evolution corresponds to the typical one relative to the macrostate M_{act}.

a past "effect") but *abductive* reasoning, by which I mean the following method of "inference to the best explanation":

$$B,\, \texttt{Typ}(B \mid A) \rightsquigarrow A \tag{10.5}$$

In other words, rather than asking what past state (or states) would be typical given the present macrostate, we should ask what past state (or states) would make our present macrostate typical. (If no such past macrostate exists, the next best explanation would be one that makes the present state *not atypical*, i.e., $\texttt{Typ}(\neg B)$ but $\neg\texttt{Typ}(\neg B \mid A)$.)

I am not able to provide a complete analysis of what makes an explanation good or best, but the following principles of "parsimony" seem reasonable:

i) If $\texttt{Typ}(B \mid A)$ and $\texttt{Typ}(B \mid A')$ but $A' \subsetneq A$, then A is (ceteris paribus) the preferable explanation. In other words, if both A and $A \cap X$ make B typical, we should infer A and not commit to an explanation that is more specific than necessary.

ii) Assuming constant ancillary conditions C, if both $\texttt{Typ}(B \mid A \cap C)$ and $\texttt{Typ}(B \mid A' \cap C)$ but $\texttt{Typ}(A \mid C)$ while $\neg\texttt{Typ}(A' \mid C)$ (A is typical but A' is not) or $\texttt{Typ}(\neg A' \mid C)$ while $\neg\texttt{Typ}(\neg A \mid C)$ (A' is *atypical* but A is not) then A is preferable to A' given C. This is the point where when we see hoof prints in the ground, we infer (based on zoological facts C that we know about the world) that they were made by a horse rather than a unicorn.

In the upshot, I submit that causal relations are defined between macrostates and in terms of typicality. In particular, causal inferences and explanations are a form of typicality inference/ explanation, notably based on *conditional* typicality. Features of the world that are typical *tout court* – not just typical *given* some prior state or event that may itself be atypical – are explained in a deeper, more conclusive sense.

In particular, such *tout court* typicality explanations are not just a fallback option when we would really like to know the causal history of the world but don't possess enough information. Once a feature of our world has been established as typical relative to all nomologically possible worlds, all explanatory pressure is relieved. To wonder further, *why* our world is typical in that respect would be irrational. Causal explanations, on the other hand, tend to lead into regress – to what caused the cause? – and since the search for a first cause or "unmoved mover" seems in vain, the explanation is never fully grounded.

Remark (Typicality facts summarizing causal facts)**.** Wilhelm (2019) argues that some typicality facts explain by summarizing part of the causal history of the world. This applies, more specifically, to typicality facts about actual ensembles that exist in our world. Wilhelm's go-to example is: "Short-tailed bobcats are typical" – which arguably summarizes part of the history of biological evolution of bobcats – and he takes this to be explanatory of the fact that Mary the bobcat has a short tail. While Wilhelm already distinguishes such typicality facts (about the actual world) from modal typicality facts (about nomologically possible worlds), I am making the further case that the latter are, in fact, what grounds causal explanations. In other words, typicality explanations are more fundamental than causal ones.

10.2 Causal and Epistemic Asymmetry

We have just said that the thermodynamic asymmetry corresponds to an asymmetry of typicality, so to speak. But then one could worry that we have applied what Price (1996) calls a "temporal double standard" in accounting for the thermodynamic arrow in the first place. The thermodynamic history of our universe is typical relative to the initial Past Hypothesis macrostate, and we took this fact to be explanatory. But the evolution is atypical with respect to the future (final?) macrostate – and haven't we insisted that explanations based on atypicality are unacceptable, that atypical facts cry out for further explanation? Indeed we have, but the atypical evolution of the universe towards the entropic past *is* explained by the low-entropy boundary condition, i.e., the *Past Hypothesis.*

There is not much more to be said here unless one is unhappy with this explanation. And there are reasons to be unhappy with it since a low-entropy initial condition is itself atypical relative to the complete phase space of the theory. In the next chapter, we will thus discuss the prospect of making do without the assumption of a special initial macrostate and establish a thermodynamic arrow as typical *tout court.* In this case, the fact that the macro-evolution in one time-direction is atypical relative to any intermediate macrostate is itself a typical phenomenon and explained by this (more basic, since unconditional) typicality fact.

In any case, it should be emphasized that if one holds a reductive view about the direction of time, the identification of the low-entropy boundary condition with the "past" is not a priori. The aim is rather to reduce the perceived differences between "past" and "future" to the thermodynamic asymmetry and/or the asymmetric boundary conditions.

An opposite point of view is advocated by Tim Maudlin (2007a), who regards the direction and, more precisely, the passage of time as metaphysically primitive. For Maudlin, it lies in the nature of time and laws that explanations go from past or initial to future or final states:

> So we have the following situation: if the asymmetrical treatment of the 'initial' and 'final' boundary conditions of the universe is a reflection of the fact that time passes from the initial to the final, then the entropy gradient, instead of explaining the direction of time, is explained by it. [...]
>
> If we are to maintain that typicality arguments have any explanatory force – and it is very hard to see how we can do without them – then there must be some account of why they work only in one temporal direction. Why are microstates, except at the initial time, always atypical with respect to backward temporal evolution? And it seems to me that we *have* such an explanation: these other microstates are *products of a certain evolution*, an evolution guaranteed (given how it *started*) to produce exactly this sort of atypicality. This sort of explanation requires that there be a fact about which states produce which. That is provided by a direction of time: earlier states produce later ones. Absent such a direction, there is no account of one global state being a cause and another an effect, and so no account of which evolutions from states should be expected to be atypical and typical in which directions. If one only gets the direction of causation from the distribution of matter in the space-time, but needs the direction of causation to distinguish when appeals to typicality are and are not acceptable, then I don't see how one could *appeal* to typicality considerations to *explain* the distribution of matter, which is what we want to do. (pp. 131-134)

I do not believe there's any sense borne out by physics in which a *microstate* $X(t_0)$ produces the states $X(t)$ for $t > t_0$ but not for $t < t_0$. There is a sense in which a *macrostate* $M(t_0)$ produces $M(t)$ for $t > t_0$ rather than $t < t_0$, namely that

$$\texttt{Typ}\,(M(t) \mid M(t_0)) \iff t \geq t_0 \tag{10.6}$$

because of the entropic arrow. This is sufficient to capture the causal intuitions that we gather from our manifest image of the world[1] but extrapolating them to the fundamental (microscopic) level is questionable. In any case, Maudlin and I seem to agree that there are deep connections between the asymmetries of entropy, typicality, and causation but disagree on issues of priority. Here, I am interested in pursuing the reductive program with respect to the arrow of time and seeing how far it can take us.

My starting point will be the question, in what sense the difference between causal and abductive inferences – which we have associated with predictions and retrodictions, respectively – could account for our experience of a direction of time. Other authors (e.g., Reichenbach (1956); Price (1996); Albert (2000); Callender (2016)) have identified two phenomena, in particular, that must be accounted for:

1. We have records of the past but not the future.
2. We can influence the future but not the past (or at least have the very strong impression that we can).

Asymmetry of records

We find a dinosaur bone in the ground and conclude that a past macrostate which would have typically evolved into the present state with a dinosaur bone is a state containing a dinosaur. The bone is thus a record of a dinosaur in the past, based on abductive reasoning in the sense of (10.5). Notably, the usual argument that "a state containing a dinosaur is much more unlikely than a random fluctuation producing a bone" has no basis in our analysis. At no point did we associate the phase space measure of a macro-region with an intrinsic probability.

We can also *predict*, for the distant future, that the bone will further decay (as almost anything else), but this might seem less interesting than a dinosaur, of course. Why do we find fossil records from animals that have existed in the past but not from animals that will exist in the future? Because dying and decaying is an entropy-increasing (thermodynamically irreversible) process that typically occurs only in one time direction.

This example is somewhat special in that it can be conceived in terms of the thermodynamic history of the dinosaur alone. In other cases, where we can consider a subsystem as isolated, the idea that its present state tells us more about the past than about the future may be simply unjustified. For instance, one could argue that a half-melted ice cube is as much a "record" of a full ice cube in the past as of a puddle of water in the future.

Many records, however, are relevant because they seem to tell us something about an interaction with other systems. When we expose a photographic film to take a picture, the final record state has lower entropy than the initial state. Still, if we find

[1] But remember that we always have to look at the full picture: a carpenter produces not only a chair but also a lot of waste and heat.

a photograph and ask what state in the past would have typically produced it, we can infer a film in a camera being exposed to light reflected from the scenery depicted in the image. ("Producing" here means nothing more than evolving into a state containing the photograph, which may then branch off as an independent subsystem.) There is no autonomous evolution of the photograph itself that would make its current state typical, hence we infer that it was part of a larger system in which it has interacted in the past.

In the first instance, when we observe a subsystem B at time t_0, we don't know with what other systems it may have interacted at other times $t \neq t_0$. However, we infer *past* interactions under the constraint that they must typically produce the present state of B, while no such constraint is justified for *future* interactions. Someone may come and throw the photograph into a fire. The evolution of a heap of ash into a photo is atypical. Hence, we can conclude that the photograph has not been burned at any time $t < t_0$, but *not* that it won't be burned at some time $t > t_0$ (because, to repeat, the thermodynamic asymmetry implies that macro-evolutions towards the entropic past *are*, in general, atypical relative to the future state).

For the same reason, I submit that, whatever its neurophysiological basis may be, *memory* must work in the same fashion. A past interaction with a system A at $t < t_0$ can make a brain state at time t_0 typical (relative to the state of $A \oplus B$ at t) but a future interaction will not. In order to create "records" of future events, the process of perception and memorization would have to be extremely sensitive to microscopic details which would come at a very high cost but with little reward from the point of view of Darwinian evolution. A system supposed to store reliable information about macroscopic events can do so only for events in the entropic past of the corresponding record state.

This analysis can be complementary to that of David Albert (2000, 2015), who argues that a record is not one instantaneous state of a system from which we infer something about an earlier state but two diachronic states – a "ready state" and a "record state" – from which we infer something about events occurring in between. If we consider, for instance, a frictionless billiard game and we know that the black ball had momentum $\boldsymbol{p}_0$ at time t_0 and momentum $\boldsymbol{p}_1 \neq \boldsymbol{p}_0$ at t_1, we can infer that it must have collided with at least one other ball at some point in the time-interval (t_0, t_1). There is no entropic arrow in this (idealized) example, assuming a frictionless billiard with perfectly elastic collisions. However, when we observe the ready state $\boldsymbol{p}_0$, we do not know the ball's momentum at t_1. When we observe the record state $\boldsymbol{p}_1$, we remember the ready state $\boldsymbol{p}_0$ at t_0 and can thus infer that a collision has occurred in the meantime.

At least in such cases, Albert's observation strikes me as perfectly correct. As an account of the epistemic asymmetry, the argument may seem circular or to lead into an infinite regress: Our records are records of the past and not the future because we know more about the past (i.e., the ready states). And we know more about the past than about the future because we have records of the past. What breaks the circle according to Albert is the Past Hypothesis, which he calls “the mother ... of all ready conditions” (Albert, 2000, p. 118). But it has never seemed quite plausible to me (and many other’s, I think) how the cosmological boundary conditions of the universe could figure in an inference about billiard balls, especially since we don’t know yet what the initial macrostate of our universe actually was. Albert’s response is that “some crude, foggy, partly unconscious, radically incomplete, but nonetheless perfectly serviceable acquaintance with the consequence of the past hypothesis and the statistical postulate and the microscopic equations of motion will very plausibly have been hard-wired into the cognitive apparatus of any well-adapted biological species” (Albert, 2015, p. 39). To me, this claim sounds itself a tiny bit foggy and incomplete. I have, therefore, proposed a different explanation why biological species adapted to the thermodynamic asymmetry would remember the past but not the future. That analysis (based on a different notion of “record”) may provide some of the missing pieces for Albert’s account – and vice versa.

Asymmetry of influences

Let us now consider the asymmetry of influences, i.e., point 2 from above. We have at least an illusion of agential control over a limited number of physical degrees of freedom; first and foremost over our body, and then, by extension, our immediate surroundings with which our body can interact (cf. Loewer (forthcoming)).[2] As limited as this (perceived) control may be, it is enough to “decide” between macro-configurations that differ radically with respect to their typical evolution into the future. Just think of David Lewis’ example of president Nixon deciding whether or not to push the atomic button (Lewis, 1979). A slight movement of the finger may cause – or not cause – a nuclear war.

Now, we have already explained why causal influences do not go from present or future states to past ones. But what about the abductive inferences (10.5) that we deemed appropriate for retrodiction: Would the best explanation, i.e., the macrostate(s) typically evolving into Nixon’s then-present state, have been different if Nixon had pressed the button? On the one hand, the answer is: Yes, of course, a different explanandum would have required a different explanans (duh). But it may seem like calling the inferred past state an “explanation” is doing too much work in dismissing this counterfactual dependence as unremarkable. Thus it seems relevant to note that, on the other hand, the past macrostates which make it typical for Nixon to push or not push the button must arguably include the physical state of Nixon’s brain – to which Nixon himself has no direct epistemic access. In general, we may introspect about the *reasons* for our decision but rarely about their physical cause.

Notably, nothing in our account undercuts the counterfactual “if Nixon had pushed

[2] I am, by the way, a compatibilist about free will, but the issue is beyond the scope of our present discussion, and I would largely endorse the eternalist view of Hoefer (2002).

the button, then the past history of the universe would have been different." This counterfactual is true, period. It just isn't constitutive of a causal influence according to our analysis, which understands the relevant counterfactual dependence in terms of typical macro-histories. It is then an observation of psychological rather than metaphysical importance that the states thus dependent on our choices (and not too far removed from the present) are largely *external* when they lie in the (entropic) *future*, and *internal* – involving, in particular, our brain state – when they lie in the (entropic) *past*. I believe that this at least begins to explain why an inference of the form (10.1) *feels* so different from an inference of the form (10.5); why the former rather than the latter would be associated with the impression – and be it only an illusion – of having an impact on the world.

Albert (2000, 2015) and Loewer (e.g., 2007a, 2012b) argue that the possible macro-histories are nomologically more constrained towards the past because they must converge in the Past Hypothesis macro-region. This way of reducing not the causal to the thermodynamic asymmetry but both to the Past Hypothesis gives the causal asymmetry more physical substance – at least if the Past Hypothesis is understood as a physical law, as they suggest.[3] I remain open to this option, which is neither presupposed nor contradicted by the analysis provided here. But I hesitate to assign such a distinguished status to the Past Hypothesis because of my hopes (to be pursued in the next chapter) that we might be able to avoid it. In any case, one could say that the Mentaculus of Albert and Loewer grounds the causal asymmetry directly in what they take to be the laws of nature (plus an analysis of counterfactuals) while in my account, it is one or two steps further removed. I am not unhappy about this, since I have long been convinced that causality is an effective and somewhat anthropomorphic concept whose legitimation from the fundamental laws of physics goes only so far.

[3] Metaphysically, Albert and Loewer are Humeans, of course, so they don't admit any fundamental causal relations, either.

Chapter 11

Arrow(s) of Time without a Past Hypothesis

11.1 The Easy and the Hard Problem of Irreversibility

What is the difference between past and future? Why do so many physical processes occur in only one time direction, despite the fact that they are governed or described, on the fundamental level, by time-symmetric microscopic laws? These questions are intimately linked to the notion of entropy and the second law of thermodynamics. From the point of view of fundamental physics, it is the second law of thermodynamics that accounts for such phenomena as that gases expand rather than contract, that glasses break but don't spontaneously reassemble, that heat flows from hotter to colder bodies, that a car slows down and doesn't accelerate once you stop hitting the gas. All these are examples of irreversible processes, associated with an increase of entropy in the relevant physical systems.

Goldstein (2001) – possibly inspired by Chalmers' discussion of the mind-body problem (Chalmers, 1995)– distinguishes between the *easy part* and the *hard part* of the problem of irreversibility. The easy part of the problem is: *Why do isolated systems in a state of low entropy typically evolve into states of higher entropy (but not the other way round)?* The answer to this question was provided by Ludwig Boltzmann, who reduced the second law of thermodynamics to the statistical mechanics of point particles. We have discussed it in some detail in Chapter 9.

The easy problem of irreversibility can be arbitrarily hard from a technical point of view if one seeks to obtain rigorous mathematical results about the convergence to equilibrium in realistic physical models. It is easy in the sense that, conceptually, Boltzmann's account is well understood and successfully applied in physics and mathematics – despite ongoing (but largely unnecessary) controversies and misconceptions, some of which we have addressed.

The hard problem begins with the question: *Why do we find systems in low-entropy states to begin with if these states are atypical?* Often the answer is that *we* prepared them, creating low-entropy subsystems for the price of increasing the entropy in their environment. But why then is the entropy of this environment so low – most strikingly in the sense that it allows *us* to exist? If one follows this rationale to the end, one comes to the conclusion that the universe as a whole is in a state of low entropy (that is, globally, in a spatial sense; we don't just find ourselves in a low-entropy pocket in an otherwise chaotic universe) and that this state must have evolved from a state of even lower entropy in the distant past. The latter assumption is necessary to avoid the absurd conclusion that our present macrostate – which includes all our memories and records of the past – is much more likely the product of a fluctuation out of equilibrium than of the low-entropy past that our memories and records actually record. In other words: only with this assumption does Boltzmann's account "make it plausible not only that the paper will be yellower and ice cubes more melted and people more aged and smoke more dispersed in the future, but that they were less so (just as our experience tells us) in the past" (Albert, 2015, p. 5). For a good discussion of this issue, see also Feynman (1967, Ch. 5) and Carroll (2010).

In sum, the hard part of the problem of irreversibility is to explain the existence of a *thermodynamic arrow of time in our universe*, given the fact that the universe is governed, on the fundamental level, by reversible microscopic laws. And the standard account today involves the postulate of a very special (since very low-entropy) initial macrostate of the universe. Albert (2000) coined for this postulate the now-famous term *Past Hypothesis* (PH). But the status of the Past Hypothesis is highly controversial. Isn't the very low-entropy beginning of the universe itself a mystery in need of scientific explanation?

11.2 The Controversy over the Past Hypothesis

In the literature, by and large three different stances have been taken towards the status of the Past Hypothesis:

1. The low-entropy beginning of the universe requires an explanation.

2. The low-entropy beginning of the universe does not require, or allow, any further explanation.

3. The Past Hypothesis is a law of nature (and therefore does not require or allow any further explanation).

The first point of view is largely motivated by the fact that our explanation of the thermodynamic arrow is based on a typicality reasoning. Assuming a low-entropy initial macrostate of the universe, Boltzmann's analysis allowed us to conclude that *typical* microstates relative to this macrostate will lead to a thermodynamic evolution of increasing entropy. On the flip side, we have argued that atypical facts are usually the kind of facts that cry out for further explanation (cf. Maudlin (2020)). And to accept the PH is precisely to assume that the initial state of our universe was atypical, relative to all possible microstates, in that it belonged to an extremely small (i.e., very low-entropy) macro-region. Penrose (1989) estimates the measure of this macro-region relative to the available phase space volume to be at most $1 : 10^{10^{123}}$ – a mind-bogglingly small number.

Arguably, the explanatory pressure is somewhat mitigated by the fact that the PH entails only a special initial macrostate rather than a microscopic fine-tuning. More precisely, the relevant boundary condition can be characterized in a simple and non-question-begging way (that is, in terms of its low entropy and without invoking the explanandum of an entropy-increasing evolution). Nonetheless, the necessity of a PH implies that our universe looks very different from a typical model of the fundamental laws of nature – and this fact alone raises legitimate concerns.

Figure 11.1: God picking out the special (low-entropy) initial conditions of our universe. Source: Penrose (1989).

The second point of view is, in particular, defended by Callender (2004a,b). While Callender is also sympathetic to the third option (regarding the PH as a law), he makes the broader case that a) there is no single feature of facts – such as being atypical – that makes them require explanation, and b) the conceivable explanations of the Past Hypothesis aren't much more satisfying than accepting it as a brute and basic fact. Notably, Ludwig Boltzmann himself eventually arrived at a similar conclusion:

> The second law of thermodynamics can be proved from the mechanical theory if one assumes that the present state of the universe, or at least that part which surrounds us, started to evolve from an improbable state and is still in a relatively improbable state. This is a reasonable assumption to make, since it enables us to explain the facts of experience, and one should not expect to be able to deduce it from anything more fundamental. (Boltzmann, 1897)

The third and final option is most prominently advocated by David Albert (2000) and Barry Loewer (2007a) in the context of the Humean *best system account* of laws. Upon their view, the laws of nature consist in

a) The microscopic dynamical laws.

b) The Past Hypothesis

c) A probability (or typicality) measure on the initial macro-region.

This package is the "Mentaculus" (Loewer, 2012b) that we have already introduced in Chapter 5. Here it should be noted, however, that the proposition which Albert wants to grant the status of a law is not that the universe started in *any* low-entropy state. The PH, in its current form, is rather a placeholder for "the macrocondition ... that the normal inferential procedures of cosmology will eventually present to us" (Albert, 2000, p. 96). Ideally (I suppose), physics will one day provide us with a nice, simple, and informative characterization of the initial macrostate of our universe – maybe something along the lines of Penrose's Weyl curvate conjecture (Penrose, 1989) – that would strike us as "law-like." But this is also what many advocates of option 1 seem to hope for as an *explanation* of the PH. So while option 3 sounds like the most clear-cut conclusion about the status of the PH, it is debatable to what extent it settles the issue. The more we have to rely on future physics to fill in the details, the less is already accomplished by calling the Past Hypothesis a law of nature. Moreover, if we had such a "law-like" characterization of the initial boundary conditions of the universe, we would still have the option to interpret them as a Humean law, or as nomologically necessary in a metaphysically more robust, i.e., anti-Humean, sense.[1] Tying option 3) to Humean metaphysics may thus be philosophically convenient but not at all necessary.

11.3 Thermodynamic Arrow without a Past Hypothesis

In recent years, Sean Carroll together with Jennifer Chen (2004; see also Carroll (2010)), and Julian Barbour together with Tim Koslowski and Flavio Mercati (2013, 2014, 2015) independently put forward audacious proposals to explain the arrow of time *without* the postulate of an atypical initial state. While Barbour's arrow of time is not, strictly speaking, an *entropic* arrow (but rather connected to a certain notion of complexity), Carroll's account is largely based on the Boltzmannian framework, although with a crucial twist. For this reason, we shall focus on the Carroll account first, before comparing it to the theory of Barbour et al. in Section 11.6.

The crucial assumption of Carroll and Chen is that the relevant measure on the state space of the universe is unbounded, allowing for macrostates of *arbitrarily high entropy* (while we shall assume that none has *infinite* entropy). Then, *every* macrostate is a non-equilibrium state from which the entropy can typically increase in both time

[1] At least I don't see why anti-Humeans must be committed to dynamical laws only, though some of the major anti-Humean positions, such as the production view of Maudlin (2007a) might be.

directions, defining a thermodynamic arrow – or rather two opposite ones – on either side of the entropy minimum. A typical entropy curve (one hopes) would thus be roughly parabolic or "U-shaped," attaining its global minimum at some moment in time and growing monotonously (modulo small fluctuations) in both directions from this vertex (Fig. 11.2). Barbour et al. (2015) describe such a profile as "one-past-two-futures," the idea being that observers on each branch of the curve would identify the direction of the entropy minimum – which the authors name the *Janus point* – as their past. In other words, we would have two future-eternal epochs making up the total history of the universe, with the respective arrows of time pointing in opposite directions.

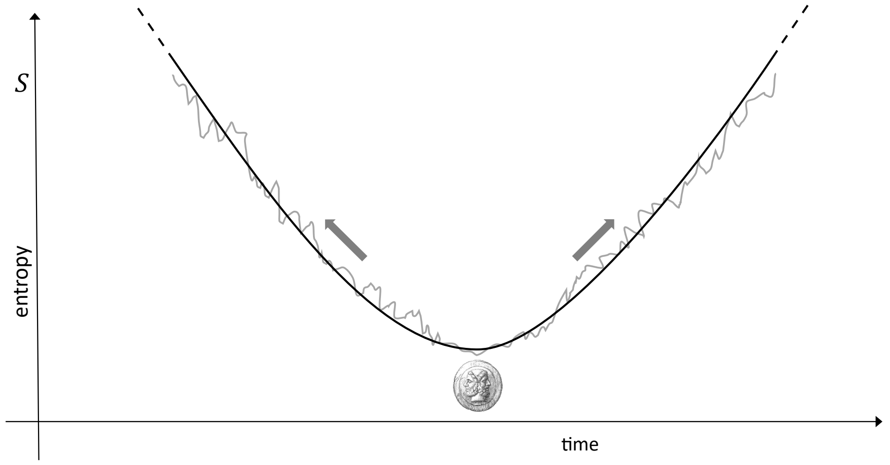

Figure 11.2: Typical entropy curve (with fluctuations and interpolated) for a Carroll universe. The arrows indicate the arrow(s) of time on both sides of the "Janus point" (entropy minimum).

The Carroll model is intriguing because it is based on the bold, yet plausible assumption that the universe has no equilibrium state – a crucial departure from the "gas in the box" paradigm that is still guiding most discussions about the thermodynamic history of our universe (cf. Barbour (2017)). And it is particularly intriguing for anybody worried about the status of the Past Hypothesis because it seeks to establish the existence of a thermodynamic arrow in the universe as *typical*. This is in notable contrast to the standard account, in which we saw that an entropy gradient is typical only under the assumption of atypical – and time-asymmetric – boundary conditions.

Prima facie, it seems plausible that an eternal universe with unbounded entropy would exhibit the U-shaped entropy profile shown in Fig. 11.2. For if we start in *any* macrostate, the usual Boltzmannian arguments seem to suggest that typical microstates in the corresponding macro-region lead to entropy-increase in both time directions (since there are always vastly larger and larger macro-regions, corresponding to higher and higher entropy values, that the microstate can evolve into). And then, any sensible regularization of the phase space measure would allow us to conclude that a U-shaped entropy profile is typical *tout court*, that is, with respect to all possible micro-histories.

However, if we assume, with Carroll, a non-normalizable measure – that assigns an infinite volume to the total phase space and thus allows for an unbounded entropy –, the details of the dynamics and the phase space partition must play a greater role than usual in the Boltzmannian account. For instance, the measure of low-entropy macro-regions could sum up to arbitrarily (even infinitely) large values, exceeding those of the high-entropy regions. Or the high-entropy macro-regions could be arbitrarily far away in phase space, so that the dynamics do not carry low-entropy configurations into high-entropy regions on relevant time scales. The first important question we should ask is therefore:

> Are there any interesting and realistic dynamics that give rise to typical macro-histories as envisioned by Carroll and Chen?

The original idea of Carroll and Chen (2004) is as fascinating as it is speculative. The authors propose a model of eternal spontaneous inflation in which new baby universes (or "pocket universes") are repeatedly branching off existing ones. The birth of a new universe would then increase the overall entropy of the multiverse, while the baby universes themselves, growing from a very specific pre-existing state (a fluctuation of the inflaton field in a patch of empty de-Sitter space), would typically start in an inflationary state that has much lower entropy than a standard big bang universe. This means, in particular, that our observed universe can look like a low-entropy universe, with an even lower-entropy beginning, even when the state of the multiverse as a whole is arbitrarily high up the entropy curve. The details of this proposal are beyond the scope of this thesis and do not (yet) include concrete dynamics or a precise definition of the entropy.

In more recent talks, Carroll discusses a simple toy model – essentially an ideal gas without a box – in which a system of N non-interacting particles can expand freely in empty space. The only macro-variable considered is the moment of inertia, $I = \sum_{i=1}^{N} \mathbf{q}_i^2$ (in the center-of-mass frame), providing a measure for the expansion of the system. It is then easy to see that I will attain a minimal value at some moment in time t_0, from which it grows to infinity in both time directions (cf. equation (11.7) below). The same will hold for the associated entropy since a macro-region, corresponding to a fixed value of I, is just a sphere of radius $\sqrt{I}$ in the position coordinates (while all momenta are constant). The entropy curve will thus have the suggested U-shape with vertex at $t = t_0$. A detailed discussion of this toy model can be found in Reichert (2012), as well as Goldstein et al. (2016).

I am not going to discuss these two models in more detail since I have little to add to the references cited above. Instead, I am going to argue in Section 11.5 that there exists a dynamical theory fitting Carroll's entropy model that is much less speculative than baby universes and much more interesting, physically, than a freely expanding system of point particles. This theory is *Newtonian gravity.* It will also allow us to draw interesting comparisons between the ideas of Carroll and Chen and those of Barbour, Koslowski, and Mercati.

First, however, we want to address the question, whether this entropy model would even succeed in explaining away the Past Hypothesis. Are typical macro-histories as envisioned by Carroll and sketched in Fig. 11.2 sufficient to ground sensible inferences about our past and future? Or would we still require, if not the PH itself, then a close variant, an equally problematic assumption about the specialness of the observed universe?

11.4 The (dispensible) Role of the Past Hypothesis

The question to be addressed in this section is thus the following:

> Can Carroll's entropy model ground sensible statistical inferences about the thermodynamic history of our universe without assuming (something akin to) a Past Hypothesis?

To approach this issue, and clarify the role of the PH in the standard account, we have to disentangle two often confounded questions:

i) Given the fundamental laws of nature, what do typical macro-histories of the universe look like? In particular: is the existence of a thermodynamic arrow typical?

ii) Given our knowledge about the present state of the universe, what can we reasonably infer about its past and future?

The answer to question i) will, in general, depend on the dynamical laws as well as cosmological considerations. If we have infinite time and a finite maximal entropy, a typical macro-history will be in thermodynamic equilibrium almost all the time, but also exhibit arbitrarily deep fluctuations into low-entropy states, leading to periods with a distinct entropy gradient, i.e., a local thermodynamic arrow. This *fluctuation scenario* was, in fact, one of Boltzmann's attempts to resolve to the hard problem of irreversibility (Boltzmann, 1896b).

However, to assume a fluctuation as the origin of our thermodynamic arrow is highly unsatisfying, Feynman (1967, p. 115) even calls it "ridiculous." The reason is that fluctuations which are just deep enough to account for our present macrostate are much more likely (i.e., would typically occur much more frequently[2]) than fluctuations producing an even lower-entropy past from which the current state could have evolved in accordance with the second law. We would thus have to conclude that we are

[2] e.g. in the sense $\limsup\limits_{T\to+\infty} \frac{1}{T}\Big(\#\text{fluctuations to entropy } S \text{ in the time-interval } [-T,T]\Big)$

currently experiencing the local entropy minimum, that our present state – including all our records and memories – is, in fact, the product of a random fluctuation rather than a lower-entropy past. Feynman makes the further case that the fluctuation scenario leads not only to absurd conclusions about the past but to wrong ones about the present state of the universe, as it compels us to assume that our current fluctuation is not any deeper than necessary to explain the evidence we already have: If we dig in the ground and find a dinosaur bone, we should not expect to find other bones nearby. If we stumble upon a book about Napoleon, we should not expect to find other books containing the same information about a guy called Napoleon. The most extreme form of this reductio ad absurdum is the *Boltzmann brain problem* (see, e.g., Carroll (2010) for a nice discussion): a fluctuation that is just deep enough to account for your empirical evidence (many people claim) would produce only your brain, floating in space, with the rest of the universe at equilibrium. You should thus conclude that this is, by far, the most likely state of the universe you currently experience.

The only possible escape in such a fluctuation scenario is to invoke the additional postulate – a form of Past Hypothesis – that the present macrostate is not the bottom of the fluctuation, but has been preceded by a sufficiently long period of entropy increase from a state of much lower entropy, still. In this context, the PH would thus serve a *self-locating* function, taking the form of an indexical proposition that locates our present state on the upwards-slope of a particularly deep fluctuation (Fig. 11.3).

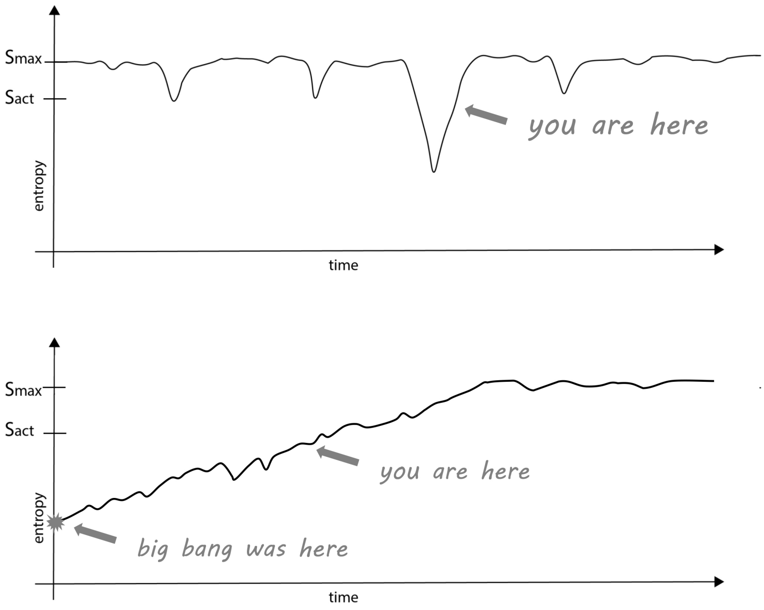

Figure 11.3: Self-location hypothesis in the fluctuation scenario (upper image) and big bang scenario (lower image) with bounded entropy. Time-scales in the upper image are much larger than below and periods of equilibrium are much longer than depicted.

The now standard account assumes a bounded entropy and a relatively young universe – about 13.8 billion years according to estimates from modern big bang cosmology. In this setting (we interpret the big bang as the actual beginning of time), a typical history would not have any thermodynamic arrow at all (the time scale of $\sim 10^{10}$ years is too short for significant entropy fluctuations on cosmological scales). Thus, we need the PH to account for the existence of a thermodynamic arrow in the first place by postulating a low-entropy boundary condition at the big bang. A self-locating proposition is still crucial and hidden in the assumption of a young universe. Winsberg (2012)) makes it explicit in what he calls the "Near Past Hypothesis" (NPH), which is that our present state lies between the low-entropy beginning of the universe and the time of first relaxation into equilibrium. Without such an assumption – and assuming that the universe is eternal in the future time direction – we would essentially be back in a fluctuation scenario with all its Boltzmann-brain-like absurdities. In a future-eternal universe with bounded entropy, there are still arbitrarily many entropy fluctuations that are just deep enough to account for our present evidence (but not much deeper). And we would still have to conclude that we are much more likely in one of these fluctuations than on the initial upwards slope originating in the very low-entropy big bang (cf. Loewer (2020)).

The self-locating role of the PH (which I take to include the NPH – for what would be the point otherwise?) is thus indispensable. And it is, in fact, the indexical proposition involved, rather than the non-dynamical boundary condition, that I would be surprised to find among the fundamental laws of nature as we consider this option for the status of the Past Hypothesis.

Sean Carroll's model postulates an eternal universe and unbounded entropy, suggesting that typical macro-histories will have the U-shaped entropy profile depicted in Fig. 11.2. If this works out – and I will argue that it does, but at least see no reason why it couldn't – the existence of a thermodynamic arrow (respectively two opposite ones) will be *typical*. (For completeness, we could also discuss the option of a temporally finite universe and unbounded entropy, but this model doesn't seem to add much of interest.)

In the upshot, Carroll's model could indeed explain the existence of a thermodynamic arrow without postulating a Past Hypothesis over and above the microscopic laws. It may still turn out that we need to invoke a PH for purposes of self-locating, *if* the theory would otherwise suggest that our present macrostate is the global entropy minimum, i.e., has not evolved from a lower-entropy past. The relevant version of the PH may then take the form of an indexical clause – stating that our present state is high up the entropy curve – or be a characterization of the entropy minimum (Janus point) of our universe. (In the first case, the PH would locate the present moment within the history of an eternal universe; in the latter, it would first and foremost locate the actual universe within the set of possible ones.) But it is not obvious why the Carroll model would lead to the conclusion that our current state is near the entropy minimum, and the issue actually belongs to our second question – how to make inferences about our past – to which we shall now turn.

Predictions and Retrodictions

The most straightforward response to question ii) – how to make inferences about the past or future – is the following method of statistical reasoning: Observe the current state of the universe (respectively a suitably isolated subsystem), restrict the pertinent probability (more correctly: typicality) measure to the corresponding macroregion in phase space, and use the conditional measure to make probabilistic inferences about the history of the system. I shall call this *naive evidential reasoning* (reviving a terminology introduced in an unpublished 2011 draft of Goldstein et al. (2016)). The negative connotation is warranted because we know that while this kind of reasoning works quite well for *predictions* – inferences about the future – it leads to absurd, if not self-refuting, conclusions when applied for *retrodictions* – i.e., inferences about the past.

The standard move to avoid this predicament is to employ the PH to block naive evidential reasoning in the time direction of the low-entropy boundary condition. For sensible retrodictions, we learn, one must conditionalize on the low-entropy initial state in addition to the observed present state. It is rarely, if ever, noted that an appeal to a PH may be sufficient but not necessary at this point. The key is to appreciate that the second question – how to make inferences about the past and future of the universe – must be addressed subsequently to the first – whether a thermodynamic arrow in the universe is typical. For if we have good reasons to believe that we live in a universe with a thermodynamic arrow of time, this fact alone is sufficient to conclude the irrationality of retrodicting by conditionalizing the phase space measure on the present macrostate.

Indeed, we have seen it in the previous chapter: The Boltzmannian analysis implies that in a system with a thermodynamic arrow, the evolution towards the future (the direction of entropy increase) looks like a *typical* one relative to any intermediate macrostate, while the actual microstate is necessarily atypical with respect to its evolution towards the entropic past. Hence, the fact that naive evidential reasoning doesn't work towards the entropic past can be inferred from the existence of a thermodynamic arrow; it does not have to be inferred from the assumption of a special initial state. The explanation of the thermodynamic arrow, in turn, may or may not require a special initial state, but this was a different issue – discussed above.

If the relevant physical theory tells us that a thermodynamic arrow is typical, i.e., exists in almost all possible universes, we have a very strong theoretical justification for believing that we actually live in a universe with a thermodynamic arrow. And if we believe that we live in a universe with a thermodynamic arrow, a rational method for making inferences about the past is not naive evidential reasoning, but the inference to the best explanation (10.5) discussed in Chapter 10. Instead of asking what past state is typical given the present macrostate (or adjust our credence in the past macrostate M_0 to $\mathbb{P}(M_0 \mid M_{act})$, if such a probability even makes sense), we should ask what past state would typically evolve into the present one, i.e., "bet" on macrostates M_0

that maximize $\mathbb{P}(M_{act} \mid M_0)$. If we find a dinosaur bone, we should infer a past state containing a dinosaur. If we find history books with information about Napoleon, we should infer a past state containing a French emperor by the name of Napoleon. In particular, considering the universe as a whole, the fact that it has evolved from a lower-entropy state in the past is *inferred*, rather than assumed, by this kind of abductive reasoning.

By now, it should be clear that the debate is not about whether the assertion of a low-entropy past is *true*, but about whether it is an *axiom.* And the upshot of our discussion is that if the existence of a thermodynamic arrow in the universe turns out to be typical, we can consider our knowledge of the low-entropy past to be reasonably grounded in empirical evidence and our best theory of the microscopic dynamics (as any knowledge about our place in the history of the universe arguably should).

Another way to phrase the above analysis goes as follows: Naive evidential reasoning applied to both time directions will always lead to the conclusion that the current macrostate is the (local) entropy minimum. However, if we know that we observe a universe (or any other system) with a thermodynamic arrow, we also know that this conclusion would be wrong *almost all the time.* More precisely, it would be wrong unless we happened to observe *a very special period* in the history of the universe in which it is close to its entropy minimum.

Goldstein, Tumulka, and Zanghì provide a mathematical analysis of this issue in the context of Carroll's toy model of freely expanding particles (Goldstein et al., 2016). Their discussion shows that the two opposing ways of reasoning – typical microstates within a given macro-region versus typical time-periods in a history characterized by a U-shaped entropy curve – come down to different ways of regularizing the unbounded phase space measure by choosing an appropriate cut-off. Goldstein et al. then argue against the first option, corresponding to naive evidential reasoning, and say that certain facts about the past amount to "pre-theoretical" knowledge. I have provided a concurrent argument based more explicitly on a theoretical (Boltzmannian) analysis. Nonetheless, from a formal point of view, a certain ambiguity remains. In Section 11.6, we will discuss how the relational framework of Barbour et al. is able to improve upon this situation.

The mystery of our low-entropy universe

Another possible objection to the Carroll model (disregarding baby universes) goes as follows: Doesn't the fact that the entropy of the universe could be arbitrarily high make its present very low value – and the even lower value at the Janus point – only more mysterious? In other words: doesn't the fact that the entropy could have been arbitrarily high only increase the explanatory pressure to account for the specialness of the observed universe?

I believe that the Carroll model precludes any *a priori* expectation of what the entropy of the universe should be. If it can be arbitrarily (but not infinitely) large, any possible value could be considered “mysteriously low” by skeptics. This is what we called a “Morgenbesser case” in Ch. 1.2 when we proposed a characterization of acceptably brute facts: Why is the entropy of our universe so low? *If it were any higher, you’d still be complaining!*

I guess divergent intuitions about this question are possible, however, and the ambiguity is once again paralleled by mathematical issues arising from the non-normalizability of the phase space measure.[3] I’ll have to leave it at that as far as the discussion of the Carroll model is concerned, exploring instead in a later section how the shape space theory of Barbour et al. is able to resolve the issue. First, though, I owe the reader some evidence that a discussion of Carroll universes is not pure speculation but that Newtonian gravity might, in fact, provide a relevant example.

11.5 Entropy of a Classical Gravitating System

There is a lot of confusion and controversy about the statistical mechanics of classical gravitating systems, despite the fact that statistical methods are commonly and successfully used in areas of astrophysics that are essentially dealing with the Newtonian N-body problem (see, e.g., Heggie and Hut (2003)). (An excellent paper clearing up much of the confusion is Wallace (2010)); see Callender (2010) for some problematic aspects of the statistical mechanics of gravitating systems, and Padmanabhan (1990) for a mathematical treatment.) Some examples of common claims are:

a) Boltzmann‘s statistical mechanics is not applicable to systems in which gravity is the dominant force.

b) The Boltzmann entropy of a classical gravitating system is ill-defined or infinite.

c) An entropy increasing evolution for a gravitating system is exactly opposite to that of an ideal gas. While the tendency of the latter is to expand into a uniform configuration, the tendency of the former is to clump into one big cluster.

I believe that the first two propositions are simply false, while the third is at least grossly oversimplified. However, rather than arguing against these claims in the abstract, I shall provide a demonstration to the contrary by proposing an analysis of a classical gravitating system in the framework of Boltzmann’s statistical mechanics (based on joint work with Paula Reichert).

We start by looking at the naive calculation, along the lines of the standard textbook computation for an ideal gas, that finds the Boltzmann entropy of the classical gravitating system to be infinite (see, e.g., Kiessling (2001)). For N gravitating parti-

[3] In particular, if we tried to interpret this measure probabilistically (what we don’t do upon the typicality view), we would run into the paradox that any finite range of entropy values has probability zero.

cles with (for simplicity equal) mass m in a volume V, we have

$$S(E,N,V) := k_B \log|\Gamma(E,N,V)| = k_B \log\Big[\frac{1}{h^{3N}N!}\int\limits_{V^N}\int\limits_{\mathbb{R}^{3N}} \delta(H-E)\ \mathrm{d}^{3N}q\,\mathrm{d}^{3N}p\Big], \tag{11.1}$$

with

$$H(q,p) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m} - \sum_{1\leq i<j\leq N} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|} \tag{11.2}$$

and

$$\int\limits_{V^N}\int\limits_{\mathbb{R}^{3N}} \delta(H-E)\ \mathrm{d}^{3N}p\ \mathrm{d}^{3N}q = C\int\limits_{V^N}\Big(E + \sum_{i<j}\frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|}\Big)^{\frac{3N-2}{2}}\ \mathrm{d}^{3N}q = +\infty. \tag{11.3}$$

For $N > 2$, the integral (11.3) diverges due to the singularity of the gravitational potential at the origin.

There is nothing mathematically wrong with the above calculation; it just doesn't actually compute what it's supposed to. One problem is that as we integrate over V^N, we sum over all possible configurations of N particles (with total energy E) within the volume V. This includes configurations in which the particles are homogeneously distributed over V, but also configurations in which most particles are concentrated in a small region of the volume (Fig. 11.4). In the case of the ideal gas in a box, the contribution of the latter is negligible since almost the entire phase space volume is concentrated on spatially homogeneous configurations. It is the entropy (or phase space volume) of this equilibrium state that we actually want to compute, and the mistake we make by including non-equilibrium configurations (in which the particles are concentrated in one half, or one quarter or one third, etc. of the volume) is so small that it is hardly ever mentioned.

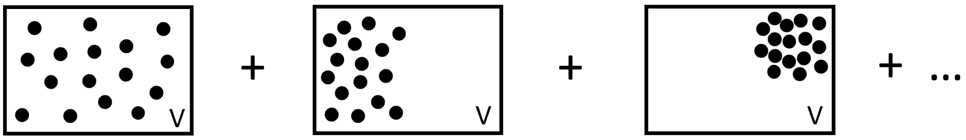

Figure 11.4: Performing the volume integral in (11.3), we sum over *all* possible configurations of the particles within the given volume V.

In the case of a gravitating system, the situation is distinctly different since the spatial configuration is correlated with the kinetic energy or, in other words, with the possible momentum configurations of the system. Simply put, for an attractive potential and constant energy, a more concentrated spatial configuration corresponds to higher kinetic energy and thus larger phase space volume in the momentum variables. The "total volume" V is thus not a good macro-variable to describe a system with gravity. In particular, if we want to know whether the entropy of a gravitating system is increasing as the configuration clusters, we have to consider macroscopic variables that actually distinguish between more and less clustered configurations.

I propose to describe a system of N gravitating point particles by the following set of macro-variables:

- $E(p,q) = \frac{p^2}{2m} + V(q) = \sum\limits_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m} - \sum\limits_{1\leq i<j\leq N} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|}$ is the total energy of the system which is a constant of motion.
- $I(q) = \sum\limits_{i=1}^{N} m(\mathbf{q}_i - \frac{1}{N}\sum\limits_{j=1}^{N} \mathbf{q}_j)^2$ is the moment of inertia that will quantify how much the particles are spread out over space. In the center of mass frame, it simplifies to $I(q) = mq^2 = \sum_{i=1}^{N} m\mathbf{q}_i^2$. Notably, we will consider particles that can expand arbitrarily in space, without any physical boundaries (like a box) confining them to a given volume. $I(q)$ can thus grow without bound.

The moment of inertia alone is still too coarse to differentiate between, let's say, a uniform configuration and a concentrated cluster with few residual particles far away. To distinguish between more and less clustered configurations, we have to introduce a further macro-variable. We choose:

- $U(q) := -V(q) = \sum\limits_{1\leq i<j\leq N} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|}$, which is just the absolute value of the potential energy. Since the total energy is $E(q,p) = T(p) + V(q)$, specifying the value of E and U is equivalent to specifying E and the kinetic energy $T(p) = \sum\limits_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m}$. An increase of $U(q)$ thus signifies both clustering and heating of the system.

  Note that defining macrostates in terms of U (respectively the potential energy) automatically takes care of the *ultraviolet divergence* in the computation of the associated entropy since the minimal particle distance r is bounded as $r \geq \frac{Gm^2}{U}$.

Evidently, I do not claim that the moment of inertia or the gravitational potential energy of the universe can be precisely measured. What makes them relevant macro-variables is, first and foremost, the fact that they are coarse-graining: many different micro-configurations of an N-particle universe realize the same values of I and U. Moreover, it will soon become clearer that these macro-variables do indeed provide relevant information about the large-scale structure of a gravitating universe.

Now, to determine the entropy of the respective macrostates, we have to compute the phase space volume corresponding to a macro-region $\Gamma(E,I,U)$, that is

$$|\Gamma(E,I,U)| = \iint\limits_{\mathbb{R}^{3N}\times\mathbb{R}^{3N}} \delta\Big(\frac{p^2}{2m} + V(q) - E\Big)\ \delta\Big(V(q) + U\Big)\ \delta\Big(mq^2 - I\Big)\ \mathrm{d}^{3N}q\,\mathrm{d}^{3N}p \tag{11.4}$$

for fixed values of E, I, and U. Unfortunately, we weren't able to solve this integral analytically (which might be, in fact, impossible). However, if we replace the sharp values of the macro-variables with a small interval $I(q) \in (I - \Delta I, I + \Delta I)$, $|V(q)| \in (U - \Delta U, U + \Delta U)$ with, e.g., $\Delta I = \frac{I}{\sqrt{N}}, \Delta U = \frac{U}{\sqrt{N}}$ (roughly a standard deviation), we can obtain the bounds:

$$C\Big(\frac{Gm^{5/2}}{\sqrt{I}U}\Big)^3 e^{-5N}(E+U)^{\frac{3N-2}{2}} I^{\frac{3N}{2}} \leq |\Gamma(E, U \pm \Delta U, I \pm \Delta I)| \leq Ce^{\sqrt{N}}(E+U)^{\frac{3N-2}{2}} I^{\frac{3N}{2}},$$

for sufficiently large values of I and U (more precisely, of the dimensionless quantity $\frac{\sqrt{I}U}{Gm^{5/2}}$) and $E \geq 0$, where C is a positive constant depending only on N and m. A precise statement and proof (valid for any E) is given in the appendix. Thus, we have

$$|\Gamma(E, U, I)| \approx const. \cdot \big(I(E + U)\big)^{\frac{3N}{2}}, \tag{11.5}$$

and, ignoring an additive constant,

$$S(E, I, U) \approx \frac{3N}{2}\Big(\log(E + U) + \log(I)\Big). \tag{11.6}$$

Typical evolutions

We now provide a discussion of this result.

1. With our choice of macro-variables, the associated Boltzmann entropy of a gravitating system is well-defined and finite. We also see that the entropy can grow without bounds, either due to continuous expansion of the system ($I \to +\infty$) and/or due to continuous clustering and self-heating ($U \to +\infty$).

2. While common wisdom says that the typical evolution of a gravitating system is one of clumping and clustering, our computation shows that clustering and expansion (as quantified by the macro-variable U and I, respectively) can contribute equally to an increase of entropy. This fits well with the observed processes of gravithermal collapse that are known to show a "core-halo" pattern (see, e.g., Heggie and Hut (2003, Ch. 23)): the configuration of masses splits into a core that collapses and heats up (increase of U) and a collection of particles on the outskirts that are blown away (increase of I).

   On even larger (cosmological) scales, a gravitating system in a homogeneous configuration can increase its entropy along both "dimensions" by forming many local clusters ("galaxies") that disperse away from each other – a process that would look very much like structure formation!

   Hence, it seems to be precisely the interplay between the opposing tendencies of clustering and expansion that makes classical gravity much more interesting, from a thermodynamic point of view, than often assumed.

3. Analytical and numerical results support the conclusion that the typical evolution of a gravitating system is one in which the entropy (11.6) increases from a minimum value in both time directions, giving rise to the U-shaped entropy curves proposed by Carroll and Chen. The first analytical result is the classical *Lagrange-Jacobi equation* for the gravitational potential:

$$\ddot{I} = 4E - 2V. \tag{11.7}$$

From this equation, which is a standard result in analytical mechanics, it follows immediately that if $E \geq 0$, the second time derivative of the moment of inertia is strictly positive (note that V is negative), meaning that $I(t)$ is a strictly convex (upwards curving) function. Together with the fact that $I \to \infty$ as $t \to \pm\infty$ (Pollard, 1967), we can conclude that the graph of I has precisely the kind of U-shape that we expect for the entropy.

Thanks to the results of Saari (1971b) and Marchal and Saari (1976), we have an even clearer picture of the asymptotic behavior of the Newtonian gravitational N-particle system. Their work studies the inter-particle distances $|\mathbf{q}_i - \mathbf{q}_j|$, as well as the dispersion from the center of mass for $t \to \infty$, independent of the total energy. They found that either the minimal particle distance goes to zero

$$\lim_{t\to\infty} r(t) := \lim_{t\to\infty} \min_{i\neq j} |\mathbf{q}_i(t) - \mathbf{q}_j(t)| = 0,$$

while the greatest particle distance goes to infinity faster than t

$$\lim_{t\to\infty} \frac{R(t)}{t} := \lim_{t\to\infty} t^{-1} \max_{i\neq j} |\mathbf{q}_i(t) - \mathbf{q}_j(t)| = \infty,$$

or the asymptotic behavior in the center-of-mass frame is characterized by

$$\mathbf{q}_i(t) = \mathbf{A}_i t + \mathcal{O}(t^{2/3}) \;\; \forall i = 1, \dots, N \quad \text{and} \quad \limsup_{t\to\infty} r > 0, \tag{11.8}$$

where $\mathbf{A}_i \in \mathbb{R}^3$ are constant vectors (possibly the zero vector). Note that since the dynamics are time-reversal invariant, the results hold for $t \to -\infty$, as well.

The first case describes so-called “super-hyperbolic escape.” This scenario is consistent with an increase of our gravitational entropy (11.6), implying both $I \to \infty$ and $U \to \infty$ as $t \to \infty$, but also includes the pathological solutions that diverge in finite time. It is the second case (when super-hyperbolic escape is excluded) in which the Newtonian N-body system is much more interesting and generally well-behaved. More precisely, we see that if (11.8) holds, all inter-particle distances fall into one of the following three classes (see Saari (1971), Cor. 1.1, together with Marchal and Saari (1976), Cor. 6):

$$|\mathbf{q}_i - \mathbf{q}_j| = L_{ij}t + \mathcal{O}(t^{2/3}), \tag{11.9}$$

$$\text{or} \qquad |\mathbf{q}_i - \mathbf{q}_j| = \mathcal{O}(t^{2/3}) \tag{11.10}$$

$$\text{or} \qquad |\mathbf{q}_i - \mathbf{q}_j| \leq L \tag{11.11}$$

asymptotically in t, with positive constants L, L_{ij}.

The result can be summarized as follows (cf. (Saari, 1971b, p. 227)): On sufficiently large time scales, the system forms clusters, consisting of particles whose mutual distances remain bounded. These clusters form subsystems (clusters of clusters) that are reasonably well isolated (energy and angular momentum are asymptotically conserved in each one of them separately), the distance between their centers of mass growing proportional to t. Finally, within each of these subsystems, the clusters separate approximately as $t^{2/3}$. In other words, the long-term behavior of such a Newtonian universe looks very much like *structure formation*, with local clumping into "galaxies" and global expansion due to galaxies and galaxy clusters receding from each other.

In regard to entropic considerations, i.e., equation (11.6), we note that the moment of inertia will grow asymptotically like $I(t) \sim t^2$, while the macro-variable $U(t)$ is at least bounded from below by some multiple of $\frac{N}{L^2}$ (assuming that the number of particles in a cluster is of order N). What happens at intermediate times? Assuming henceforth non-negative total energy, we already know that $I(t)$ is strictly convex. Together with its quadratic growth for $t \to \pm\infty$, we can conclude that it has a unique global minimum, let's say at $t = \tau$, from which it increases in both time directions. $U(t)$ will in general fluctuate, but if we exclude particle collisions and "near particle collisions" (very close encounters), it will remain bounded and not vary too quickly ($\dot{U}$ remains bounded, as well). Hence, one would expect that the graph of $(E+U(t))I(t)$ (the logarithm of which is proportional to our gravitational entropy) looks qualitatively like that of $I(t)$, namely by and large parabolic. Indeed, numerical simulations by Barbour et al. (2013, 2015) for the $E = 0$ universe (with $N = 1000$ and random initial data) support the claim that the evolution of $I \cdot U$ is well interpolated by a parabola of the form $\alpha(t-\tau)^2 + \beta$ with $\alpha, \beta > 0$. All this suggests the desired U-shaped evolution of the entropy $S(E, I, U) \approx \frac{3N}{2}\Big(\log(E+U) + \log(I)\Big)$ as a function of time for a Newtonian gravitating universe with non-negative energy. (Actually, on large time scales, the shape looks less like a U and more like $\curlyvee$ – how some children draw birds on the horizon – since $S(t)$ grows only logarithmically as $|t - \tau| \to \infty$.)

We conclude that a Newtonian gravitating universe is indeed a "Carroll universe" which has no equilibrium state and for which entropy increase (in opposite directions from a global minimum) is typical. Quite astonishingly, this entropy increase is consistent with structure formation, showing that the colloquial understanding of entropy as a "measure of disorder" does not always provide the right intuition. And contrary to another popular belief, the typical evolution of a gravitating system does not just lead to one big boring clump of matter, either.

11.6 Gravity and Typicality from a Relational Point of View

Starting from Machian / Leibnizian principles, Barbour, Koslowski, and Mercati (2013, 2014, 2015) discuss the Newtonian gravitational system from a relationalist perspective. According to the relational framework that Julian Barbour has championed over the past decades, all physical degrees of freedom are described on *shape space* S which is obtained from Newtonian configuration space by factoring out global rotations, translations, and scale, leaving us with a $3N-7$ dimensional space for an N-particle system. The configuration of N point particles is then characterized by the angles and ratios between their (Euclidean) distance vectors – or, in other words, by its *shape* – independent of extrinsic scales. The lowest-dimensional (non-trivial) shape space is that of $N = 3$ particles. In this case, the shapes are those of triangles – specified by 2 angles or the ratios between 3 distances – and the topology of shape space is that of a 2-dimensional (projective) sphere.

Considering standard Newtonian gravity on absolute space and trying to extract, so to speak, its relational essence, we have to eliminate all dependencies on extrinsic spatio-temporal structures. To this end, we restrict ourselves to models with vanishing total momentum, $\mathbf{P} = \sum_{i=1}^{N} \mathbf{p}_i \equiv 0$, and angular momentum, $\mathbf{L} = \sum_{i=1}^{N} \mathbf{q}_i \times \mathbf{p}_i \equiv 0$, excluding rotating universes and propagations of the center of mass, respectively.[4] Furthermore, the rejection of absolute time scales leads to considering only universes with zero total energy ($E \equiv 0$), since this is the only value invariant under a rescaling of time-units.

The difficult issue when it comes to formulating Newtonian gravity on shape space is the lack of scale-invariance. Newtonian gravity has models that do not rotate ($L \equiv 0$) and models that do not propagate ($P \equiv 0$) but it does not have models that do not expand ($D := \frac{1}{2}\dot{I} = \sum_{i=1}^{N} \mathbf{q}_i \cdot \mathbf{p}_i \equiv 0$; Barbour calls D the *dilatational momentum*). The characteristic size of an N-particle system is given by $\sigma = \sqrt{I}$, where $I = \sum\limits_{i=1}^{N} \mathbf{q}_i^2$ is the center-of-mass moment of inertia, and we have already seen that I can never be constant for non-negative energy (equation (11.7)) but is roughly parabolic as a function of time. In other words: an N-particle universe interacting by Newton's law

[4] Arbitrary solutions of Newtonian mechanics can be projected onto shape space, but the total angular momentum (let's say) cannot be captured by relational initial data. It corresponds to a particular choice of reference frame ("gauge"), when the shape space theory is lifted to absolute phase space; cf. Dürr et al. (2019). $\mathbf{L} = 0$ is then the only canonical choice, and the only one suggested by Machian principles.

of gravity always changes in size. In fact, the general Lagrange-Jacobi identity shows that $E = 0$ and $I \equiv const.$ is possible only if the potential is homogeneous of degree -2. This had motivated the alternative, scale-invariant theory of gravitation proposed in Barbour (2003). Here, we discuss the relational formulation of Newtonian gravity with the familiar $\frac{1}{r}$-potential.

Of course, we can (and will) insist that a "change in size" is meaningless from a relational point of view, but the process has nonetheless dynamical (and thus empirical) consequences in Newtonian theory. Simply put: for constant energy, a gravitating system that expands is also slowing down. Newtonian laws (with a potential homogeneous of degree k) are always invariant under a global rescaling of distance, $q \to \alpha q$ for constant $\alpha > 0$, when compensated by a corresponding change in time-units, $t \to \alpha^{1-k/2}t$. However, the characteristic scale $\sigma(t) = \sqrt{I(t)}$ of a gravitating system changes in time. (Compare: absolute rotations are meaningless from a relational point of view; but while the laws are invariant under time-independent rotations, particles in a rotating universe experience a centrifugal force affecting their motions.) Hence, if we eliminate this scale "by hand", namely by a time-dependent coordinate transformation $q \to \frac{q}{\sigma(t)}$, the resulting dynamics can be formulated on shape space, but will no longer have the standard Newtonian form. Instead, the dynamics become non-autonomous (time-dependent) with scale acting essentially like friction (Barbour et al., 2014).

How to capture this time-dependence without reference to external time? Barbour et al. make use of the fact that the dilatational momentum $D = \frac{1}{2}\dot{I}$ is monotonically increasing ($\ddot{I} > 0$ by equation (11.7)) and can thus be used as an internal time-parameter, a kind of universal clock. In particular, we observe that $D = 0$ precisely when I reaches its global minimum. This *central time* thus marks the mid-point between a period of contraction and a period of expansion, or better (though this remains to be justified): the Janus point between two periods of expansion with respect to opposite arrows of time. It provides, in particular, a natural reference point for parametrizing solutions of the shape space theory in terms of *mid-point data* on the shape phase space T^*S.[5]

There is one last redundancy from the relational point of view that Barbour et al. (2015) call *dynamical similarity.* It comes from the invariance of the equations of motion under a simultaneous rescaling of internal time D and shape momenta. More simply put: two solution trajectories are physically identical if they correspond to the same geometric curve in shape space, the same sequence of shapes, even if that curve is run through at different "speeds." Thus, factoring out the absolute magnitude of the shape momenta at central time, we reduce the relevant phase space (that parametrizes solutions) by one further dimension. The resulting space PT^*S (mathematically, this is the projective cotangent bundle of shape space S) is *compact*, which means, in

[5] Mathematically, this is the cotangent bundle of shape space S, just as Hamiltonian phase space is the cotangent bundle of Newtonian configuration space.

particular, that it has a *finite total volume* according to the uniform volume measure. And this is where the relational formulation, i.e., the elimination of absolute degrees of freedom, really starts to pay off. Since the uniform measure on PT^*S – which Barbour et al. take to be the natural typicality measure – is normalizable, it allows for a statistical analysis that avoids the ambiguities resulting from the infinite phase space measure in the Carroll model. Unfortunately, the construction of the measure is not entirely canonical but involves the choice of a metric on shape space. And while the choice made by Barbour and collaborators seems natural enough, the justification of the typicality measure for the shape space theory remains a critical step that would require a more in-depth analysis. (Dürr, Goldstein, and Zanghì (2019) provide an insightful discussion, though focussing on the quantum case.) Deviating from the notation of Barbour et al., we denote their measure on the reduced phase space by μ_ε.

Shape complexity and the gravitational arrow

To describe the macro-evolution of a gravitating system on shape space, Barbour and collaborators introduce a dimensionless (scale-invariant) macro-variable C_S which they call *shape complexity*:

$$C_S = -V \cdot \sqrt{I}. \tag{11.12}$$

Comparison with (11.5) and (11.6) (setting $E = 0$ and remembering that $U = -V$) shows an relationship between this shape complexity and the gravitational entropy that we computed on absolute phase space: $S(E = 0, I, U) \propto N \log(\sqrt{I} C_S)$. Recalling the previous discussion (or noting that $C_S \approx R/r$, where R is the largest and r the smallest distance between particles), we also see that low shape complexity corresponds to dense (on the scale of r) homogeneous states in absolute space, while high shape-complexity indicates "structure" – dilute configurations of multiple clusters.

On shape space, considering the simplest case of $N = 3$ particles, the configuration of minimal shape complexity is the equilateral triangle, while the configuration of maximal shape complexity corresponds to "binary coincidences" in which the distance between two particles – relative to their distance to the third – is zero. This is to say that 3-particle configurations with high shape complexity will, in general, contain a Kepler pair (a gravitational bound state of two particles) with the third particle escaping to infinity.

Above, we discussed the typical evolution of $-V \cdot I$ and found it to be roughly parabolic or U-shaped. Analogously, one can conclude that the evolution of $C_S = -V \cdot \sqrt{I}$ (in Newtonian time) will typically exhibit a V-shaped profile: it has a global minimum at central time ($D = 0$), from which it grows roughly linearly (modulo fluctuations) in both time directions (see Fig. 11.6). In the terminology of Barbour, Koslowski, and Mercati, this defines two opposite *gravitational arrows of time* with the Janus point as their common past. Note that these are not *entropic* arrows, though our previous discussion strongly suggests that the evolution of the shape complexity on shape space will align with the evolution of the gravitational entropy (11.6) on absolute space.

A remarkable feature of the relational theory, however, is that it reveals the origin of the gravitational arrow to be *dynamical* rather than *statistical*: the negative of the shape complexity corresponds to the potential that generates the gravitational dynamics on shape space. There is thus a dynamical tendency towards higher values of C_S (lower values of the shape potential). In contrast, Boltzmannian statistical mechanics suggest that entropy increase is typical because there are a great many more high-entropy than low-entropy configurations that a system could evolve into. It does not suggest that physical forces are somehow driving the system towards higher entropy states.

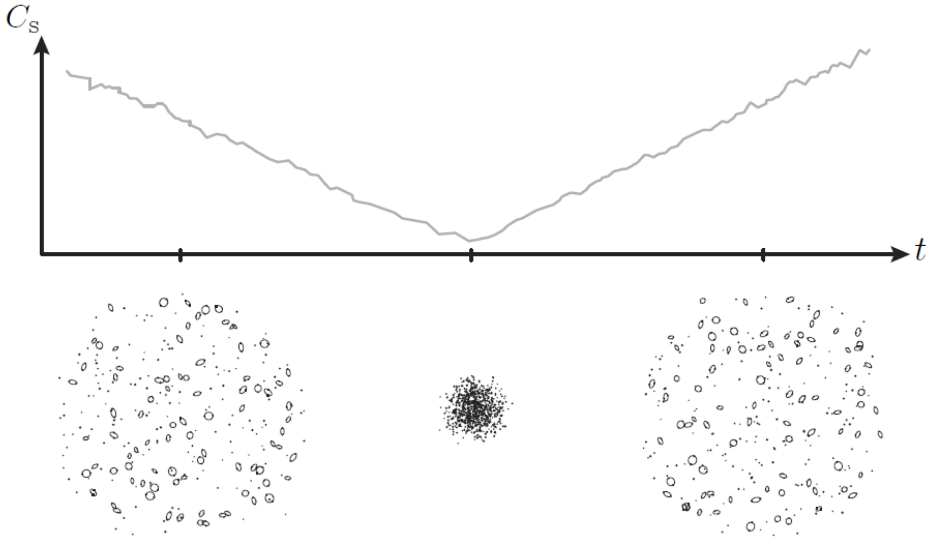

Figure 11.5: Top: evolution of the shape complexity C_S found by numerical simulation for $N = 1000$ and Gaussian initial data. Bottom: schematic depiction (not found by numerical simulation) of three corresponding configurations on Newtonian space-time. Source: Barbour et al. (2015).

Turning to the statistical analysis of the shape space theory, we are interested in determining typical values of C_S at the Janus point. To this end, we consider the measure assigned to mid-point data (Janus point configurations) with low shape complexity

$$C_S \in [C_{min}, \alpha \cdot C_{min}] := I_1, \tag{11.13}$$

respectively high shape complexity

$$C_S \in (\alpha \cdot C_{min}, \infty) := I_\infty. \tag{11.14}$$

Here, $1 < \alpha \ll \infty$ is some positive constant, and C_{min} is the smallest possible value of C_S. The key result (not yet rigorously proven but strongly substantiated by the 3-particle case and numerical experiments for large N) is that already for small values of α ($\alpha < 2$ for large N)

$$\frac{\mu_\varepsilon(I_\infty)}{\mu_\varepsilon(PT^*S)} \approx 0, \tag{11.15}$$

and consequently

$$\frac{\mu_\varepsilon(I_1)}{\mu_\varepsilon(PT^*S)} \approx 1. \tag{11.16}$$

This means: it is typical that a universe at the Janus point (the beginning of our macro-history) is in a very homogeneous state!

Regardless of the philosophical merits of relationalism, the relational theory of Barbour, Koslowski, and Mercati thus comes with two great virtues: First, it provides a sensible *normalizable* measure on the set of possible micro-evolutions that still establishes an arrow of time as typical. Even more spectacularly, typical evolutions with respect to this measure go through very homogeneous configurations at their Janus point (∼ the "big bang"). In other words, initial states that have very low entropy from the absolutist point of view come out as typical in the shape space description – provided one accepts the proposed measure on PT^*S as a natural typicality measure. This would resolve the two potential issues that we have identified in the Carroll model: the "mysteriously" low entropy of our universe, and the justification for locating our present state reasonably far away from the entropy minimum.

Entaxy and Entropy

On the other hand, Barbour et al. introduce another concept called (instantaneous) *entaxy* that I find much less compelling. The instantaneous entaxy (the authors also use the term *solution entaxy* for the measure μ_ε on PT^*S) is supposed to be the measure of a set of shape configurations corresponding to a given value of shape complexity. It thus seems *prima facie* analogous to the Boltzmann entropy defined in terms of the macro-variable C_S, with the notable exception that it *decreases* in time as the shape complexity increases. Recall, however, that the measure μ_ε was only defined on mid-point data, by cutting through the set of solution trajectories at their Janus points, so to speak. Barbour et al. now extend it to arbitrary (internal) times by stipulating that the entaxy associated with a particular value of shape complexity at *any* point in history is the measure of *mid-point* configurations with that same shape complexity.

This definition seems somewhat *ad hoc* and corresponds to comparing macrostates at different times in terms of a measure that is *not stationary* under the dynamics: A set of mid-point data will have a bigger size than the set of time-evolved configurations (phase space volume gets lost, so to speak). Indeed, on the 3-particle shape space, one can explicitly show that the points of maximal shape complexity are dynamical attractors; hence, a stationary continuous measure on shape phase space does not exist. In general, it is not even clear if stationary measures are a meaningful concept in relational mechanics since there is no absolute (metaphysical) sense in which configurations on different solution trajectories with the same internal time are *simultaneous*.

They merely happen to agree on whatever part or feature of the configuration plays the role of an internal "clock." For all these reasons, the entaxy should not be regarded as a shape analogon of the Boltzmann entropy (which is always defined in terms of a stationary measure). In particular, the fact that the gravitational arrows point in the direction of decreasing rather than increasing entaxy is not in contradiction with Boltzmannian arguments.

Finally, one may wonder whether we could compute on absolute phase space the Boltzmann entropy associated to the shape complexity or other scale-invariant macro-variables. Note that for $E = 0$, our gravitational entropy (11.6) is a function of $-VI$. Couldn't we have just computed an entropy for the macro-variable $C_S = -V\sqrt{I}$, instead? Interestingly, the answer is negative, and the reason is the following simple result:

Proposition 11.6.1. *Let μ a measure on $\mathbb{R}^n$ (equipped with its Borel sigma-algebra), which is homogeneous of degree d, i.e., $\mu(\lambda A) = \lambda^d \mu(A)$ for any measurable $A \subset \mathbb{R}^n$ and all $\lambda > 0$. Let $F : \mathbb{R}^n \to \mathbb{R}^m$ be a measurable function homogeneous of degree k, i.e., $F(\lambda x) = \lambda^k F(x)$, $\forall x \in \mathbb{R}^n$. Then we have for any measureable value-set $J \subset \mathbb{R}^m$:*

$$\mu\left(\{x \mid F(x) \in \lambda^k J\}\right) = \lambda^d \mu\Big(\{x \mid F(x) \in J\}\Big). \tag{11.17}$$

Proof.

$$\mu\left(\{x \mid F(x) \in \lambda^k J\}\right) = \mu\left(\{\lambda x \mid F(x) \in J\}\right) = \lambda^d \mu\left(\{x \mid F(x) \in J\}\right).$$

□

From this, we can immediately conclude:

Corollary 11.6.2. *If the measure μ is homogeneous of degree $d \neq 0$ and F is homogeneous of degree 0 (i.e., scale-invariant), then*

$$\mu\left(F^{-1}(J)\right) \in \{0, +\infty\}. \tag{11.18}$$

Proof. Applying (11.17) with $k = 0$ and $d \neq 0$ yields $\mu\left(F^{-1}(J)\right) = \lambda^d \mu\left(F^{-1}(J)\right)$ for any $\lambda > 0$. □

Hence, using a homogeneous phase space measure – such as the Liouville measure on $\Gamma \cong \mathbb{R}^{6N}$ – macro-regions defined in terms of scale-invariant macro-variables must have measure zero or infinity, so that the corresponding Boltzmann entropy would be ill-defined. This suggests that the concept of entropy is intimately linked to absolute scales and thus not manifestly relational. Note, in particular, that *expansion* and *heating* – processes that are paradigmatic for entropy increase (especially, but not exclusively in our analysis of gravitating systems) – require absolutes scales of distance and velocity, respectively.

This also emphasizes once again that the gravitational arrow of Barbour et al. is not an entropic arrow, although it matches – maybe accidentally, maybe for reasons I don't quite understand – the entropic arrow that we identified on absolute phase space. The result also leaves the relationalist with the following interesting dilemma: Either the notion of entropy is meaningful only for subsystems – for which the environment provides extrinsic scales – or we have to explain why the entropy of the universe is a useful and important concept *despite* the fact that it is related to degrees of freedom that are strictly speaking unphysical, corresponding to mere gauge in the shape space theory.

Can we dispense with the Past Hypothesis?

In conclusion, the works of Carroll and Chen as well as Barbour, Koslowski, and Mercati show that it is possible to establish an arrow of time as typical, without the need to postulate special boundary conditions or any other form of Past Hypothesis. By proposing the definition of a Boltzmann entropy for a classical gravitating universe, we argued that Newtonian gravity provides a relevant realization of Carroll's entropy model which can be compared to the shape space formulation of Barbour et al. We found, in particular, that the gravitational arrows identified by Barbour and collaborators in terms of shape complexity will match the entropic arrows in the theory on absolute space. The extension to other microscopic theories (and/or macroscopic state functions) will require further research.

The relationalist and the absolutist approaches both have the resources to avoid the reversibility paradox and ground sensible inferences about the past and future of the universe. However, while certain ambiguities remain in the Carroll model – resulting, in particular, from the non-normalizability of the phase space measure – those issues are resolved by the shape space theory of Barbour et al. (provided we accept their proposed typicality measure). In any case, for a Newtonian gravitating universe, Barbour's analysis suggests that homogeneous configurations at the "big bang" (Janus point) are *typical*, explaining why the universe started in what looks like a very low-entropy state from an absolutist perspective. However, if the shape space theory is actually fundamental, the "entropy of the universe" turns out to be a somewhat spurious concept whose status remains to be discussed in more detail.

Appendix: Computation of the gravitational entropy

We compute the phase space volume of the macro-region $\Gamma(E, I \pm \epsilon I, U \pm \epsilon U)$, i.e.:

$$\frac{1}{N!h^{3N}} \int d^{3N}p \int d^{3N}q \; \delta(H(\mathbf{q},\mathbf{p}) - E)$$
$$\mathbb{1}\Big\{(1-\epsilon)U \leq \sum_{\substack{i<j\\ i,j=1}} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|} \leq (1+\epsilon)U\Big\}\mathbb{1}\Big\{(1-\epsilon)I \leq \sum_{i=1}^{N} m\mathbf{q}_i^2 \leq (1+\epsilon)I\Big\}, \tag{11.19}$$

with the Hamiltonian

$$H(q,p) = \sum_{i=1}^{N} \frac{\mathbf{p}_i^2}{2m} - \sum_{1\leq i<j\leq N} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|}.$$

We shall prove the following

Proposition. If the scale-invariant quantity $\sqrt{I}U$ is large enough that

$$\sqrt{I}U \geq 4Gm^{5/2}\log(N)N^{5/2}\,, \tag{11.20}$$

we obtain the bounds:

$$|\Gamma(E, I \pm \epsilon I, U \pm \epsilon U)| \geq C\, e^{-\frac{9}{2}N}\Big(\frac{Gm^2\epsilon}{U}\Big)^3 (E + (1-\epsilon)U)^{\frac{3N-2}{2}} \Big(\frac{I}{m}\Big)^{\frac{3N-3}{2}}\,,$$
$$|\Gamma(E, I \pm \epsilon I, U \pm \epsilon U)| \leq C\,(E + (1+\epsilon)U)^{\frac{3N-2}{2}} \Big((1+\epsilon)\frac{I}{m}\Big)^{\frac{3N}{2}}\,,$$

for any $1 > \epsilon > \frac{2}{N}$ and $N \geq 4$, where $C = \frac{(2m)^{\frac{3N-2}{2}}}{2(N!)h^{3N}}\left(\Omega^{3N-1}\right)^2$, with Ω^{3N-1} the surface area of the $(3N-1)$-dimensional unit sphere.

For non-negative E, this can be simplified further by using $(E + (1+\epsilon)U)^n \leq (1+\epsilon)^n(E+U)^n \leq e^{\epsilon n}(E+U)^n$, respectively $(E + (1-\epsilon)U)^n \geq (1-\epsilon)^n(E+U)^n \geq e^{-2\epsilon n}(E+U)^n$, for $\epsilon < \frac{1}{2}$.

Proof. We first perform the integral over the momentum variables and are left with

$$\frac{(2m)^{\frac{3N-2}{2}}}{2N!h^{3N}}\,\Omega^{3N-1} \int \mathrm{d}^{3N}q \left(E + \sum_{i<j} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|}\right)^{\frac{3N-2}{2}}$$
$$\mathbb{1}\Big\{(1-\epsilon)U \leq \sum_{\substack{i<j\\ i,j=1}} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|} \leq (1+\epsilon)U\Big\}\mathbb{1}\Big\{(1-\epsilon)I \leq \sum_{i=1}^{N} m\mathbf{q}_i^2 \leq (1+\epsilon)I\Big\}.$$

From this, it is straightforward to obtain the upper bound:

$$\begin{aligned}(11.19) &\leq \frac{(2m)^{\frac{3N-2}{2}}}{2N!h^{3N}}\,\Omega^{3N-1}\frac{2\Omega^{3N-1}}{3N}(E+(1+\epsilon)U)^{\frac{3N-2}{2}}\Big[\Big((1+\epsilon)\frac{I}{m}\Big)^{\frac{3N}{2}} - ((1-\epsilon)\frac{I}{m})^{\frac{3N}{2}}\Big]\\ &\leq \frac{C}{3N}\,(1+\epsilon)^{\frac{3N}{2}}\,(E+(1+\epsilon)U)^{\frac{3N-2}{2}}\Big(\frac{I}{m}\Big)^{\frac{3N}{2}}.\end{aligned}$$

For the lower bound, we consider the set $\mathcal{B} := B_1 \times \ldots \times B_N \subset \mathbb{R}^{3N}$ defined by

$$B_j := \big\{|\mathbf{q}_j| \in [(2j-2)\xi, (2j-1)\xi]\big\},$$

with $\xi = \xi(I,N)$ to be determined soon. That is, we consider a series of concentric spheres around the origin, their radii being an increasing multiple of ξ, and configurations for which the volume between two spheres is alternately empty or occupied by a single particle. In $\mathcal{B}$, we have $mq^2 \in [I_+ - \Delta I, I_+]$ with

$$I_+ = \sum_{i=2}^{N} m\mathbf{q}_i^2 \leq m\sum_{i=2}^{N}(2i-1)^2\xi^2 = \frac{m}{3}N(4N^2-1)\xi^2 \tag{11.21}$$

and

$$\Delta I = m\sum_{i=1}^{N}[(2i-1)^2-(2i-2)^2]\xi^2 = m\sum_{i=1}^{N}[4i-3]\xi^2 \leq 2mN^2\xi^2 \leq \frac{2}{N}I_+.$$

We thus set

$$\xi := \sqrt{\frac{3I}{mN(4N^2-1)}}, \tag{11.22}$$

so that $I_+ = I$, and note that $\frac{2}{N}I < \epsilon I$ for $\epsilon > \frac{2}{N}$, so configurations in $\mathcal{B}$ are within the right range of values for the moment-of-inertia macrovariable.

Now we have to make sure to consider configurations whose potential energy is also in the right range $|V(q)| \in [U \pm \epsilon U]$.

To this end, we note that for $q \in \mathcal{B}$, the distance between two particles is bounded from below by $|\mathbf{q}_i - \mathbf{q}_j| \geq (2(j-i)-1)\xi$, and for each $1 \leq k \leq N-1$, there exists less than N particle pairs with $j-i=k$. Hence, the potential energy is bounded by

$$\sum_{i<j}\frac{Gm^2}{|\mathbf{q}_i-\mathbf{q}_j|} \leq N\sum_{k=1}^{N-1}\frac{Gm^2}{(2k-1)\xi} \leq 2N\log(N)\frac{Gm^2}{\xi} < U, \tag{11.23}$$

where we used the assumption (11.20) and the estimate

$$\sum_{k=1}^{N-1}\frac{1}{(2k-1)} \leq 1 + \sum_{k=2}^{N-1}\frac{1}{k} \leq 1 + \int_1^N \frac{1}{x}\mathrm{d}x = 1+\log(N) \leq 2\log(N), \text{ for } N \geq 4.$$

In particular, we know that for $q \in \mathcal{B}$ and, e.g., $\mathbf{q}_1 = \mathbf{0}$, we have $|V(q)| < U$. But also that $\lim_{\mathbf{q}_1\to\mathbf{q}_2}|V(q)| = +\infty$. Hence, by the mean value theorem, there exists for any

$(\mathbf{q}_2, \ldots, \mathbf{q}_N) \in B_2 \times \ldots \times B_N$ a $\lambda \in (0,1)$ such that $-V(\lambda \mathbf{q}_2, \mathbf{q}_2, \ldots, \mathbf{q}_N) = U$.

Now it is not difficult to check that if we place $\mathbf{q}_1$ at a distance of not more than $r := \epsilon \frac{Gm^2}{2U}$ from $\lambda \mathbf{q}_2$, the potential energy $|V(q)| = \sum\limits_{i=2}^{N} \frac{Gm^2}{|\boldsymbol{q}_1 - \boldsymbol{q}_i|} + \sum\limits_{1 \neq i < j \leq N} \frac{Gm^2}{|\boldsymbol{q}_i - \boldsymbol{q}_j|}$ will change by less than $||V(q)| - U| \leq \epsilon U$. Moreover, while $\mathbf{q}_1$ may no longer be in B_1, the moment of inertia $m \sum_{i=1}^N \mathbf{q}_i^2$ increases by less than $m(3\xi)^2 < \epsilon I$ and thus remains within the interval $[I \pm \epsilon I]$.

We denote by $K_r[q]$ the ball of radius r around $\lambda \mathbf{q}_2$, with $\lambda = \lambda(\mathbf{q}_2, \ldots, \mathbf{q}_N)$ as introduced above. Its volume is $\frac{4\pi}{3}\left(\frac{Gm^2\epsilon}{2U}\right)^3$. The volume of the set $B_j, j \in \{2, \ldots, N\}$ is

$$|B_j| = \frac{4\pi}{3}\left[(2j-1)^3 - (2j-2)^3\right]\xi^3 = \frac{4\pi}{3}\left[j(12j-18)+7\right] \geq \frac{16\pi}{3} j^2 \,.$$

Hence:

$$|B_2 \times \ldots \times B_N| = \prod_{j=2}^{N} |B_j| = \Big(\frac{16\pi}{3}\Big)^{N-1} \xi^{3(N-1)} \prod_{j=2}^{N} j^2 = \Big(\frac{16\pi}{3}\Big)^{N-1} \xi^{3(N-1)} (N!)^2 \,.$$

To simplify, we note that $\xi^{3(N-1)} \geq \left(\frac{3}{4}\right)^{\frac{3(N-1)}{2}} \left(\frac{I}{m}\right)^{\frac{3(N-1)}{2}} N^{-\frac{9(N-1)}{2}}$. And comparing with the area of the unit sphere, $\Omega^{3N-1} = \frac{2\pi^{3N/2}}{\Gamma\left(\frac{3N}{2}\right)}$, we use $\left(\frac{16\pi}{3}\right)^{N-1}\left(\frac{3}{4}\right)^{\frac{3(N-1)}{2}} > \Omega^{3N-1}\Gamma\left(\frac{3N}{2}\right)$. Thus:

$$\begin{aligned}(11.19) \geq & \frac{(2m)^{\frac{3N-2}{2}}}{2N! h^{3N}} \, \Omega^{3N-1} \int\limits_{K_{r(q)} \times B_2 \times \ldots \times B_N} \mathrm{d}^{3N} q \left(E + \sum_{i<j} \frac{Gm^2}{|\mathbf{q}_i - \mathbf{q}_j|}\right)^{\frac{3N-2}{2}} \mathbb{1}\Big\{\ldots\Big\}\mathbb{1}\Big\{\ldots\Big\} \\ \geq & \, C\, N^{-\frac{9(N-1)}{2}} (N!)^2 \, \Gamma\left(\frac{3N}{2}\right) \Big(\frac{Gm^2\epsilon}{2U}\Big)^3 \; (E + (1-\epsilon)U)^{\frac{3N-2}{2}} \Big(\frac{I}{m}\Big)^{\frac{3N-3}{2}} \,.\end{aligned}$$

Summing over all possible permutations of the particles over the rings in the definition of $\mathcal{B}$, we get an additional factor of $N!$. With the Sterling approximation $\Gamma(n+1) = n! > \sqrt{2\pi n}\left(\frac{n}{e}\right)^n$, we finally obtain a lower bound of the form

$$(11.19) \geq C\, e^{-\frac{9}{2}N} \Big(\frac{Gm^2\epsilon}{U}\Big)^3 (E + (1-\epsilon)U)^{\frac{3N-2}{2}} \Big(\frac{I}{m}\Big)^{\frac{3N-3}{2}} \,.$$

□

Chapter 12

Quantum Mechanics

12.1 The Measurement Problem

If, after more than a century, people are still debating Schrödinger's cat, then not just because the brilliant physicist found such a vivid image to illustrate his paradox. Rather, Schrödinger hit the nail right on its head. He formulated the *measurement problem of quantum mechanics*, and this measurement problem shows why the orthodox view of the theory is not just unsatisfactory but completely untenable.

In a nutshell, the problem with quantum mechanics is the following: There is only one equation and one physical entity defining the theory – the Schrödinger equation and the wave function – and they do not describe the phenomena as we perceive them. There are various ways to see that, for instance as follows: Suppose a system is described by a linear combination of wave functions φ_1 and φ_2 and an apparatus can display either "φ_1" or "φ_2" by interacting with the system.

In principle, this apparatus must also have a quantum mechanical description. After all, we conceive the measuring apparatus as consisting of atoms and molecules, and if all these atoms and molecules are described by a wave function, then this wave function must also provide a quantum mechanical description of the apparatus as a whole. This means that the apparatus has states Ψ_1 und Ψ_2 – pointer positions "1" and "2" corresponding to wave packets with disjoint support in configuration space – and a ready state Ψ_0, such that:

$$\varphi_i \Psi_0 \overset{\text{Schrödinger evolution}}{\longrightarrow} \varphi_i \Psi_i \ . \tag{12.1}$$

The Schrödinger time evolution (12.1), however, is linear. Therefore, a system wave function

$$\varphi = c_1\varphi_1 + c_2\varphi_2, \qquad c_1, c_2 \in \mathbb{C}, \qquad |c_1|^2 + |c_2|^2 = 1,$$

leads to

$$\varphi\Psi_0 = (c_1\varphi_1 + c_2\varphi_2)\Psi_0 \overset{\text{Schrödinger evolution}}{\longrightarrow} c_1\varphi_1\Psi_1 + c_2\varphi_2\Psi_2. \tag{12.2}$$

This seems like an absurd result. The superposition

$$c_1\varphi_1\Psi_1 + c_2\varphi_2\Psi_2 \tag{12.3}$$

describes an entangled state between system and apparatus in which the pointer position is "1" and "2" at the same time. "Pointer" here is just a stand-in for whatever registers or indicates the measurement result. In Schrödinger's famous cat experiment, φ_1 would describe a decayed atom (leading to the release of hydrocyanic acid) and φ_2 the not yet decayed atom (so that the flask of poison remains intact). (12.3) thus describes a superposition of dead cat and alive cat.

Thus, if we insist that the result of the measurement is either "1" or "2", but not both at the same time, we have the following situation: Either the wave function of the system after measurement is not (12.3). Then the Schrödinger equation is not correct, at least not always. Or the wave function of the system is indeed (12.3), but this wave function does not provide a complete description of the physical situation. In this case, we are missing precisely those variables that make the difference between a pointer pointing to the left and a pointer pointing to the right – the variables that make the difference between a dead cat and a living cat.

The orthodox answer

In order to maintain the completeness of the quantum mechanical description, John von Neumann, Werner Heisenberg, and others introduced an additional axiom into the theory. In the process of "measurement" or "observation," they postulated, the Schrödinger time evolution is suspended and replaced by a random dynamic which reduces the superposition (12.3) with probability $|c_i|^2$ to the wave function $\varphi_i\Psi_i$. However, in contrast to the Schrödinger evolution, this new dynamic, the *collapse of the wave function*, was not supposed to be described by a precise mathematical law. It is introduced *ad hoc*, as a property of "the observer." Indeed, it is precisely for this reason that the observer assumes a central role in the theory, as the subject whose act of measurement or observation brings about the physical facts. Wolfgang Pauli described this as an "act of creation lying outside the laws of nature." However, it is not even these esoteric notes of the Copenhagen school that prevent us from accepting the collapse postulate as part of a serious physical theory, but simply its "unprofessional vagueness" (Bell). We now have two contradictory dynamics for the wave function, so when exactly does one or the other apply? When exactly is a physical process considered a "measurement"? And what distinguishes an "observer" from a molecule, or a cat, or the pointer of an apparatus? Here is Bell in his brilliant article *Against measurement*:

> It would seem that the theory is exclusively concerned about 'results of measurement', and has nothing to say about anything else. What exactly qualifies some physical systems to play the role of 'measurer'? Was the

> wavefunction of the world waiting to jump [collapse] for thousands of millions of years until a single-celled living creature appeared? Or did it have to wait a little longer, for some better qualified system ... with a Ph.D.? 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? Do we not have jumping then all the time? (Bell, 2004, p. 216)

Three solutions to the measurement problem

Following Tim Maudlin (1995), a precise and general formulation of the measurement problem can be given as follows: There are three principles that a naive reading of quantum theory seems to suggest, and these three principles together are logically inconsistent; they cannot all be true.

1) The wave function of a system provides a complete description of its physical state.

2) The time evolution of the wave function always follows a linear (Schrödinger) equation.

3) Measurements (usually) have unique outcomes.

The contradiction was derived above and is, in essence, Schrödinger's cat paradox. If the wave function follows a linear time evolution, the measurement procedure (12.2) results in a macroscopic superposition (12.3). If (12.3) provides a complete description of the physical state of the measurement device, the outcome of the measurement cannot be unique. Thus: $1) \vee 2) \Rightarrow \neg 3)$. The three assumptions are mutually inconsistent, and any coherent formulation of quantum mechanics must negate at least one of them.

The negation of 1) leads to Bohmian mechanics, which postulates an ontology of point particles so that the state of a system is given by the actual particle configuration in addition to the wave function. For an N-particle system, this is a pair (ψ, Q), with $Q \in \mathbb{R}^{3N}$ the particle configuration in physical space. The role of the wave function ψ is first and foremost to guide the motion of the particles. This is expressed by a precise mathematical law in which the wave function enters. The measurement problem is solved because every system has a well-defined spatial configuration at all times, given by the positions of its constituent particles. The wave function of a measurement device (for instance) may be in a superposition (12.3), but the actual configuration Q describes a pointer pointing *either* to the left *or* to the right.

The negation of 2) leads to objective collapse theories like GRW (after Ghirardi, Rimini, and Weber (1986)), which modify the Schrödinger equation by a stochastic (non-linear) collapse term. This collapse law is such that systems with few degrees of freedom hardly ever collapse, while macroscopic superpositions are typically destroyed almost instantly. The crucial point is that the collapse of the wave function is now described by a precise mathematical law which is valid at all times; it is not, as in orthodox quantum mechanics, an ad hoc postulate tied to vague notions like "measurements" or "observers."

The negation of 3) leads to the Many Worlds theory. If we insist on 1) and 2), we have no other choice but to accept macroscopic superpositions such as those of "dead cat" and "alive cat" as a consequence of quantum mechanics. The radical conclusion, generally attributed to Hugh Everett III., is that both parts of this superposition describe a real physical state. But wouldn't we then observe two cats rather than one? Or maybe a cat in an absurd hybrid state of "dead" and "alive"? No, we would not, because the superposition of the wave function doesn't stop with the cat. It would come to include the experimentalist herself, the laboratory, indeed the entire universe, so that everything joins in the splitting described by (12.3). In the last resort, this description of nature thus comprises two "worlds," corresponding to the decoherent branches of the wave function: In one world, the radioactive atom has decayed, the cat has been poisoned to death, and the experimentalist is sad. In the other, the atom has not decayed, the cat is alive and well, and the experimentalist takes the animal back home. Because of the linearity of the Schrödinger equation, the two "copies" of the experimentalist can never interact, and decoherence makes it practically impossible to bring the dead cat and the living cat back into interference. Thus, both worlds can exist in parallel, without observers in one ever (directly) perceiving the other. They have a common past but are causally disjoint with respect to their future evolution.

It is not my goal here to argue that Bohm, GRW, and Everett provide the only solutions to the measurement problem, but by and large, the three strategies exhaust the space of logical possibilities. Either we admit additional physical variables over and above the wave function, or we modify the linear wave equation, or we accept a Many-Worlds picture.

Connecting the wave function to the world

From an orthodox perspective, the measurement problem is often phrased as the question when (and how) the collapse of the wave function occurs, i.e., when exactly the linear Schrödinger evolution is suspended in favor of the random state reduction. The deeper issue behind the measurement problem, however, is how the wave function (or the quantum formalism more generally) is supposed to connect at all to the world that we experience. The deeper issue, in other words, is not what to make of a superposed wave function of "dead cat" and "alive cat" but what a complex-valued function on a high-dimensional configuration space (or maybe some vector in an abstract Hilbert space) has to do with a cat in the first place.

Before even getting at the infamous collapse postulate, standard quantum mechanics tries to relate the wave function to observable quantities by other abstract axioms involving self-adjoint operators and their eigenvalues. If we look at an honest textbook like Cohen-Tannoudji et al. (1991), we find postulates such as:

1. At a fixed time t_0, the state of a physical system is defined by specifying a wave-function $\psi(x, t_0)$.

2. Every measurable physical quantity Q is described by an operator $\hat{Q}$; this operator is called an observable.

3. The only possible result of the measurement of a physical quantity Q is one of the eigenvalues of the corresponding observable $\hat{Q}$.

4. When the physical quantity Q is measured on a system in the normalized state ψ, the probability $\mathbb{P}(q_n)$ of obtaining the non-degenerate eigenvalue q_n of the corresponding observable $\hat{Q}$ is
$$\mathbb{P}(q_n) = \int \phi_n^* \psi \tag{12.4}$$
where ϕ_n is the normalized eigenvector of $\hat{Q}$ associated with the eigenvalue q_n.

So the “unprofessionally vague” notions of “measurement” (or “measurable,” or “measured”) appear already in postulates 2–4 before we arrive at the contradiction between wave function collapse (postulate 5) and the linear Schrödinger evolution (postulate 6). In particular, there is no clear prescription for which observable-operator corresponds to which “physical quantity,” or what measurement procedure is suitable for measuring it (and no way to tell by analyzing the textbook theory). We rather have to trust that the theoretician will know what computations to do, and the experimentalist will know what experiment to perform so they can compare their data. Vagueness and ambiguity is certainly unacceptable when it comes to the dynamical laws of a theory, but why should they be more acceptable when it comes to how the formalism of the theory relates to the physical and empirical facts?

In order to provide a precise and objective description of nature, modern quantum theories have, by and large, followed two different strategies.

One is the *primitive ontology* program (see, e.g., Allori et al. (2008, 2014); Esfeld (2014a) and Bell’s notion of “local beables” in Bell (2004, Ch. 7)) which admits additional physical variables, over and above the wave function, that represent the fundamental constituents of matter in space and time. Such theories – with Bohmian mechanics as the prime example – thus relieve the wave function from the burden of representing matter, its role being instead a dynamical one for the evolution of the primitive ontology.

The other program can be subsumed under the umbrella-term of *wave function functionalism.* It tries to develop an objective description of nature in terms of the wave function alone by locating macro-objects as patterns in the wave function (e.g., Albert (2013); Ney (2015)). We will discuss this attempt below in the context of the Many Worlds theory.

In any case, if the dynamical laws are mathematically consistent and the ontology of a theory clear – if it is clear, in other words, how the theory seeks to describe matter in space and time – there cannot be any ambiguities or contradictions. The description may be wrong or empirically adequate, but it cannot be paradoxical.

Objective collapse theories are usually regarded as a third approach to solving the measurement problem but actually fall in either one of the two camps just described. The original GRW theory (now sometimes called GRW0) is a theory about the wave function alone. However, there is still a difference between a cat and the wave function of a cat – even a collapsed one – and GRW0 faces the same challenges as the Many Worlds theory in making the connection. Nowadays, it is thus common to equip the GRW theory with a primitive ontology, as well. One proposal (due to John Bell (2004, Ch. 22)) regards the collapse centers themselves as the primitive ontology – discrete "matter flashes" in space and time that constitute macro-objects (GRWf). Another proposal (due to GianCarlo Ghirardi et al. (1995)) uses the GRW wave function to define a continuous mass density field in physical space or spacetime (GRWm). Either way, one can make the case that the stochastic collapse law adds nothing to the Bohmian (or Everettian) solution of the measurement problem (Esfeld, 2018). However, since it leads to certain predictions that differ from those of the unitary quantum theories, it may just turn out to be empirically more correct.

12.2 Born's Rule and the Measurement Process

A common first reaction to the measurement problem is to insist that the wave function was never meant to describe the actual physical facts but that only its statistical interpretation according to Born's rule is significant. A superposition like (12.3) should thus be read as saying that the measurement outcome is "1" with probability $|c_1|^2$ *or* "2" with probability $|c_2|^2$. It is this statistical law, after all, that is experimentally confirmed with great precision. Fair enough, but merely pointing to the Born rule does not solve the measurement problem. The Schrödinger equation is a deterministic equation and according to this equation, the wave function at the end of the experiment is always (12.3). If this wave function provides a complete description of system and apparatus, the outcome of the measurement is always the same. The identical physical state, if complete, cannot on some occasions describe a measurement apparatus whose pointer points to the left and on other occasions a measuring apparatus whose pointer points to the right (cf. Maudlin (1995)). So again, either the linear Schrödinger equation is incorrect, or we are missing precisely those physical variables whose probability distribution the Born rule is supposed to describe.

That said, let us apply the Born rule to our measurement scenario and see where it takes us. We describe the configuration space of the complete system by coordinates

$\boldsymbol{q} = (\boldsymbol{x}, \boldsymbol{y})$, where $\boldsymbol{x} \in \mathbb{R}^k$ are the coordinates of the measured system (S) and $\boldsymbol{y} \in \mathbb{R}^m$ those of the measurement device (D). According to Born's rule, we have

$$\begin{aligned}
\mathbb{P}(\text{pointer points to 1}) &= \int_{\operatorname{supp}\Psi_1} |c_1\varphi_1\Psi_1 + c_2\varphi_2\Psi_2|^2 \, \mathrm{d}^k x \, \mathrm{d}^m y \qquad (12.5)\\
&= |c_1|^2 \int_{\operatorname{supp}\Psi_1} |\varphi_1\Psi_1|^2 \mathrm{d}^k x \, \mathrm{d}^m y \\
&+ |c_2|^2 \int_{\operatorname{supp}\Psi_1} |\varphi_2\Psi_2|^2 \mathrm{d}^k x \, \mathrm{d}^m y \\
&\quad +2\mathrm{Re}\Big(c_1 c_2 \int_{\operatorname{supp}\Psi_1} (\varphi_1\Psi_1)^* \varphi_2\Psi_2 \mathrm{d}^k x \, \mathrm{d}^m y\Big) \qquad (12.6)\\
&\approx |c_1|^2 \int |\varphi_1\Psi_1|^2 \mathrm{d}^k x \, \mathrm{d}^m y = |c_1|^2 \; . \qquad (12.7)
\end{aligned}$$

Here, one has to note that since the supports of the two pointer wave functions on configuration space are disjoint (or nearly so), the real part (12.6) is zero (or nearly so). The probability of the outcome "1" is thus $|c_1|^2$, and the probability of the outcome "2" is $|c_2|^2$, just as the rules of textbook quantum mechanics suggest. But what exactly have we calculated here?

The response of Bohmian mechanics corresponds to the common way of speaking: Born's rule provides the probability distribution of the particle positions constituting the device with its pointer. $|c_1|^2$ is thus the probability that, at the end of the measurement process, the pointer points left, indicating the measurement result "1."

In the GRW theory, the same computation has a different meaning. Here, Born's rule provides first and foremost a probability distribution for the center of collapse. Unless the collapse centers themselves are interpreted as the ontology of the theory (the matter *flashes* mentioned above) $|c_1|^2$ is thus, first and foremost, the probability that (12.3) collapses onto a wave function localized in the support of Ψ_1.

The interpretation of Born's rule in the Many Worlds theory is difficult. Here, it doesn't make sense to say that $|c_1|^2$ is the probability that the measurement outcome "1" occurs because *all* possible outcomes occur with certainty. We shall, therefore, postpone the issue to discuss probabilities in the Many Worlds theory in more detail.

Finally, one may ask what the meaning of Born's rule is according to the orthodox (Copenhagen) quantum theory. One could say that the above computation describes a "position measurement" of the pointer. Then, $|c_1|^2$ is the probability that the pointer points left, to "1," if we look at it, but decidedly *not* the probability that the pointer points left even if nobody looks. Alternatively, one could forbid the computation altogether and insist that the probabilities must come from an observable-operator (maybe a "cat-aliveness-operator"). The most orthodox answer of all – Bohr's answer – is that the computation is forbidden because a measurement device is too big to have a wave function. If all this doesn't sound too serious, then because, indeed, it cannot be taken seriously.

Observable operators as statistical book-keepers

In textbook quantum mechanics, much ado is made about the *observable-operators* that are supposed to represent (somehow) observable quantities with their eigenvalues corresponding to the possible measurement values. Thus, I want to discuss very briefly how the observable operators arise as nothing but convenient "book-keepers" for measurement statistics.[1] For further details, I refer the reader to Dürr et al. (2004) (reprinted as Ch. 3 in Dürr et al. (2013a)) and Ch. 7 in Dürr and Lazarovici (2020).

Basically, it suffices to consider equations (12.1 – 12.7). Coupling of a system to a measurement device leads to a canalization of the wave function into decoherent (orthogonal) branches corresponding to different pointer states:

$$\varphi\Psi_0 = \left(\sum_i c_i\varphi_i\right)\Psi_0 \longrightarrow \sum_i c_i\varphi_i\Psi_i.$$

Now let P_i be the orthogonal projection onto φ_i. In Dirac notation $P_i = |\varphi_i\rangle\langle\varphi_i|$. Then, we immediately obtain:

$$\langle\varphi, P_i\varphi\rangle = |c_i|^2, \tag{12.8}$$

corresponding to the Born probability for the pointer position i (i.e., $Y \in \operatorname{supp}\Psi_i$) as computed in (12.7). If, for any $i \geq 1$, the measurement value indicated by the pointer position i is $\alpha_i \in \mathbb{R}$, the expectation value of the result of the measurement is:

$$\sum_{i\geq 1}\alpha_i|c_i|^2 = \langle\varphi, \sum_{i\geq 1}\alpha_i P_i\varphi\rangle = \langle\varphi, \hat{A}\varphi\rangle \tag{12.9}$$

with the self-adjoint operator

$$\hat{A} = \sum_{i\geq 1}\alpha_i P_i. \tag{12.10}$$

Mathematically, the right-hand-side of (12.10) is the *spectral decomposition* of $\hat{A}$.

Two points are important to take away. First and foremost, that the Born rule for "position measurements" is sufficient to ground the entire measurement formalism of quantum mechanics. Second, that the "observable values" arise, in general, only through the process of measurement and the canalization of the wave function into macroscopically disjoint branches, which can then be associated with the eigenspaces of some linear operator on Hilbert space. All the confusion about "quantum logic", "hidden variables", "metaphysical indeterminism" etc. comes only if one tries to think of quantum observables as fundamental and of their eigenvalues as intrinsic properties that a system might possess prior to, or independent of, the measurement process (cf. Daumer et al. (1996); Bell (2004); Lazarovici et al. (2018); Dürr and Lazarovici (2020)).

[1] An understanding that one could have already gathered from von Neumann's seminal *Mathematische Grundlagen der Quantenmechanik* (1932) had the operators not developed a "life of their own" in the Heisenbergian tradition.

Example (Spin Measurement)**.** An instructive example is the measurement of spin on a spin-1/2-particle. If one sends a spinor wave function $\psi_0 = \phi_0(\boldsymbol{x})\binom{\alpha}{\beta},\ |\alpha|^2 + |\beta|^2 = 1$ (in the z-spin eigenbasis) through a Stern-Gerlach-magnet oriented in z-direction, the wave function splits into spatially separating parts

$$\psi_t = \phi_+(\boldsymbol{x}, t)\begin{pmatrix}\alpha\\0\end{pmatrix} + \phi_-(\boldsymbol{x}, t)\begin{pmatrix}0\\\beta\end{pmatrix}.$$

ϕ_+ is deflected in positive z-direction and ϕ_- in negative z-direction, and the experiment is such that after a sufficient amount of time, the two wave parts will no longer overlap. At least in Bohmian mechanics, it now makes sense to ask which of the two wave packets is guiding the particle. In any case, the probability for *spin UP* and *spin DOWN* is simply the probability that the particle is found in the support of ϕ_+ (above the symmetry axis) respectively ϕ_- (below the symmetry axis) when registered by a detector or on a photographic plate. Using Born's rule, we compute

$$\mathbb{P}(\text{Spin UP}) = \mathbb{P}(\boldsymbol{X} \in \operatorname{supp}\phi_+) = |\alpha|^2 \int |\phi_+(\boldsymbol{x}, \mathrm{t})|^2\, \mathrm{d}^3\mathrm{x} = |\alpha|^2,$$
$$\mathbb{P}(\text{Spin DOWN}) = \mathbb{P}(\boldsymbol{X} \in \operatorname{supp}\phi_-) = |\beta|^2 \int |\phi_-(\boldsymbol{x}, \mathrm{t})|^2\, \mathrm{d}^3\mathrm{x} = |\beta|^2.$$

These probabilities can be read off easily from the projections to the spin components $\binom{1}{0}$ respectively $\binom{0}{1}$. Written in matrix, we have

$$P_+ = \begin{pmatrix}1 & 0\\0 & 0\end{pmatrix},\ P_- = \begin{pmatrix}0 & 0\\0 & 1\end{pmatrix},$$

and immediately obtain

$$\langle\psi, P_+\psi\rangle = |\alpha|^2,\ \langle\psi, P_-\psi\rangle = |\beta|^2. \tag{12.11}$$

The expectation value is, accordingly,

$$\begin{aligned}\frac{\hbar}{2}\,\mathbb{P}(\text{spin UP}) - \frac{\hbar}{2}\,\mathbb{P}(\text{spin DOWN}) &= \frac{\hbar}{2}\langle\psi, P_+\psi\rangle - \frac{\hbar}{2}\langle\psi, P_-\psi\rangle\\ &= \langle\psi, \frac{\hbar}{2}(P_+ - P_-)\psi\rangle = \langle\psi, \frac{\hbar}{2}\sigma_z\psi\rangle.\end{aligned}$$

Here, the Pauli matrix $\frac{\hbar}{2}\sigma_z$, commonly called the "z-spin-observable", appears as the book-keeping operator associated with the experiment.

Typicality and Observation

It is instructive to carry our computation (12.5) one step further, and a consider a "measurement of the pointer position" by another system C. We may think of an "observer" looking at the measurement device, although I would prefer a camera or some other system under no suspicion of consciousness. The spatial resolution of such an observation could be finer than the spread of the "pointer states" Φ_i[2], but we shall consider the simplest case in which the measurement interaction leads to a wave function of the form

$$c_1\varphi_1\Phi_1\Psi_1 + c_2\varphi_2\Phi_2\Psi_2, \tag{12.12}$$

where Ψ_1 is concentrated on a region $\mathcal{L}$ of the configuration space of C corresponding to the camera recording a pointer pointing left, and Ψ_2 is concentrated on a region $\mathcal{R}$ corresponding to the camera recording a pointer pointing right. (And where it is assumed that the recorded measurement device has been perfectly isolated up to this point, so there is no environmental decoherence.)

We will now analyze the situation in Bohmian terms – which are the simplest ones –, but the transfer at least to collapse theories is straightforward. So, what is the probability that the pointer actually points to the left, i.e., $Y \in \mathrm{L} := \operatorname{supp}\Phi_1$, while the camera records a pointer pointing right, i.e., $Z \in \mathcal{R} := \operatorname{supp}\Psi_2$? We find

$$\mathbb{P}(Y \in \mathrm{L}, Z \in \mathcal{R}) = \int_{\mathbb{R}^k \times \mathrm{L} \times \mathcal{R}} |c_1\varphi_1\Phi_1\Psi_1 + c_2\varphi_2\Phi_2\Psi_2|^2 \,\mathrm{d}^k x\,\mathrm{d}^m y\,\mathrm{d}^n z \approx 0, \tag{12.13}$$

since Φ_2 is (approximately) zero on L, while Ψ_1 is (approximately) zero on $\mathcal{R}$, hence both $\Phi_1\Psi_1$ and $\Phi_2\Psi_2$ are just about zero on $\mathrm{L} \times \mathcal{R}$. Simply put: if you look where the pointer is, you will *typically* see the pointer where it is.

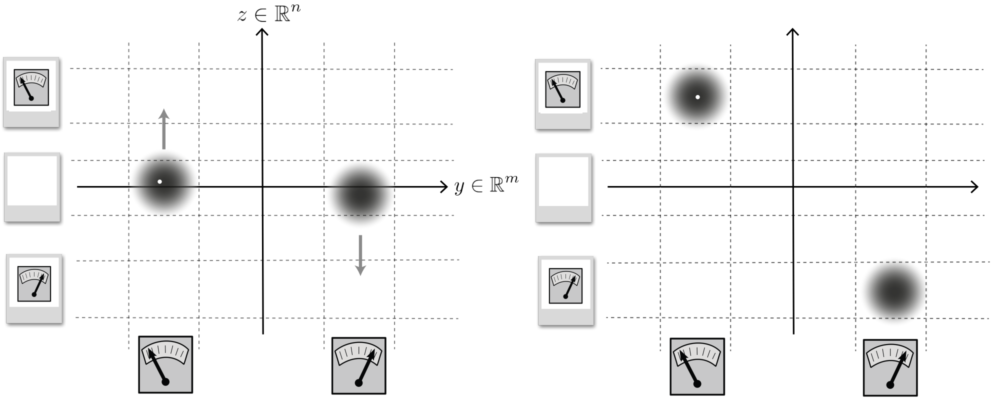

Figure 12.1: Sketch of a position measurement (recording of the pointer position) in configuration space. On the z-axis, the degrees of freedom of the device, on the y-axis, those of the camera. The dot indicates the actual configuration of the system $D \otimes C$.

[2] Thus corresponding to a Schrödinger evolution $\Phi_i \longrightarrow \sum_j \Phi_{ij}\Psi_j$, where $\sum_j \Phi_{ij} = \Phi_i$ and the Ψ_j are the record states of the "observer."

However, it is quite realistic to assume that the wave packets Φ_i or Ψ_i have long "tails" – i.e., some overlap in configuration space – so that $\mathbb{P}(Y \in \mathrm{L}, Z \in \mathcal{R})$ is not exactly zero but only nearly so (as indicated by the $\approx$ sign in (12.13)). Hence, there is a very small, yet non-zero probability that the pointer configuration points to the left (at least for a short period of time), while the camera – or "observer" – sees a pointer pointing to the right. If the measurement devices are somewhat accurate, this probability will be so small as to be practically negligible, but the atypical outcome is still *possible* according to the theory. Would this mean that the configuration Y of Bohmian particles (or GRW flashes) does not correspond to the "real" pointer position? No, it means precisely what the theory says, namely that there is an extremely small, yet non-zero probability that the pointer points left, while the camera records a pointer pointing right.

And this shouldn't be all that surprising upon reflection. Also according to classical electrodynamics, it is possible, yet *atypical*, that I see the moon to my right although it is actually to my left, because what I see is a very special fluctuation in the electromagnetic field. It is also possible, yet atypical, that I hold a thermometer – or my finger – in hot water but register a very low temperature because all the fast particles happen to move away from it.

The quantum mechanical (or Bohmian) analysis of the measurement process is thus a nice illustration of the fact that atypicality can always undermine the reliability of observations. Consequently, any inference from empirical evidence has to rely on a form of Cournot's principle, viz., on the assumption that the evidence has not been produced by an atypical or very low probability event.

12.3 Quantum Equilibrium

We have already talked about the derivation of Born's rule in Bohmian mechanics that goes back to Dürr, Goldstein, and Zanghì (1992): The Bohmian laws make Born statistics typical. For nearly all possible initial configurations of the universe – with respect to the unique equivariant measure induced by the universal wave function – ensembles of subsystems with effective wave function φ are $|\varphi|^2$-distributed. Quantum statistics thus hold in *quantum equilibrium.*

A mathematically nice, yet didactically unfortunate fact is that the typicality measure on the configuration of the universe and the typical distribution of subsystems take the same functional form in terms of the respective wave functions. Occasionally, this gives rise to the misguided criticism that the Bohmian derivation of the Born rule is "circular": $|\psi|^2$ in, $|\psi|^2$ out. It is thus important to keep in mind that the universal wave function Ψ on the one hand and the effective wave function φ of subsystems on the other have a different mathematical, physical, and metaphysical status. In particular, the $|\varphi|^2$-distribution in terms of effective wave functions is indeed derived, not assumed; quite analogous to the way in which the Maxwell distribution in classical statistical mechanics comes out as the equilibrium distribution of an ideal gas.

A vocal critic of the typicality account is Antony Valentini, who proposed instead a "quantum H-theorem" showing the convergence of a non-equilibrium distribution to the Born distribution (Valentini, 1991a,b; Valentini and Westman, 2005). Valentini's criticism is largely based on the just-mentioned "circularity objection" and an insistance that typicality is synonymous with probability (which it is not) (Valentini, 1996, 2020). His dissenting opinion on what an explanation should look like is, however, worth addressing in brief. For Valentini, the fact that quantum equilibrium is typical with respect to the natural measure doesn't justify the conclusion that our universe has been in equilibrium all along; he wants to know how it got there.

Throughout this thesis, we have worried about the Past Hypothesis and the question why our universe is not in thermodynamic equilibrium. Valentini's worry here seems to be quite the opposite, why a Bohmian[3] universe should be in quantum equilibrium and, more specifically, that if it did start out in quantum equilibrium, there could no dynamical account for the fact. Three remarks seem pertinent here:

1. The typicality explanation is based on the Bohmian dynamics. It is the dynamics that distinguish the typicality measure as the unique equivariant measure and make the Born statistics typical. (For the relevance of equivariance, see Ch. 5.5.)

2. Bohmian ensembles in non-equilibrium converge to the Born distribution for the same reason that classical gases converge to a Maxwell distribution: the vast majority of possible micro-configurations realize the respective equilibrium distribution. This is, of course, exactly what equilibrium means in Boltzmann's sense.

3. As with classical systems, not *all* non-equilibrium configurations in Bohmian mechanics converge to equilibrium; only most of them do. Thus, somewhat ironically, Valentini's account would also require typicality, although conditioned on an atypical initial macro-region that he assumes without need. Valentini responds that the "exceptional" initial conditions are ruled out empirically "by the observation of equilibrium today" (Valentini, 2020). But then we could always conclude, based on the observation of some phenomenon, that the initial conditions were such as to realize the phenomenon (unless the phenomenon is strictly impossible according to the theory). If all Valentini claims to have proven is that convergence to quantum equilibrium is *not impossible*, the result is very weak.

In effect, this entire thesis makes the case that Valentini has it backward. Atypical initial conditions, i.e., quantum non-equilibrium, would be puzzling and put the theory in question, while the fact that Born statistics come out as typical *tout court* in

[3] Valentini insists on the name *de Broglie-Bohm pilot wave theory* as opposed to *Bohmian mechanics*, which was mainly established through the work of Dürr, Goldstein, and Zanghì (DGZ). While it would be an exaggeration to speak of two different theories, Valentini's "interpretation" differs in several respects from that of DGZ.

Bohmian mechanics is the best case come true. This is not to say that the "quantum H-theorem" is completely without merit, but that it is mostly academic. In contrast to Boltzmann's H-theorem, the set of cases to which it actually applies in our universe is, for all we know, the empty set. Finally, Valentini (2020) claims that quantum equilibrium makes the Bohmian theory "unfalsifiable," by which he really means: empirically equivalent to standard quantum mechanics. Again, the opposite is true. Quantum non-equilibrium could give rise to any correlation – or lack of correlation – between the states of observed systems and the outcomes of measurements on those systems. In other words, allowing for atypical initial conditions would make the theory unfalsifiable because atypical initial conditions could produce virtually any data we like.

Thermodynamic arrow in Bohmian mechanics

We have just drawn some parallels (and contrasts) between quantum equilibrium and thermodynamic (non-)equilibrium. There is indeed an important question that needs to be addressed: If our universe, conceived as a Bohmian universe, is already in quantum equilibrium, where does the thermodynamic arrow come from? The received view to date is that it comes from the universal wave function, which started out in a special, i.e., low-entropy, macro-region of Hilbert space. To make this somewhat precise, we shall briefly summarize the generalization of Boltzmann's statistical mechanics to quantum states, according to Goldstein et al. (2010):

We can consider a (normalized) wave function Ψ as a microstate in a Hilbert space $\mathcal{H}$. More precisely, we shall restrict the system to a finite-dimensional "energy shell," corresponding to a degenerate eigenvalue of the Hamiltonian. Furthermore, we consider a partition

$$\mathcal{H} = \bigoplus_{\alpha \in \mathcal{A}} \mathcal{H}_\alpha \tag{12.14}$$

of the Hilbert space (energy shell) into orthogonal subspaces of varying though finite dimension (maybe determined by a set of relevant observables). These subspaces correspond to the Boltzmannian macro-regions and their respective quantum Boltzmann entropy is

$$S(\mathcal{H}_\alpha) = k_B \log\left(\dim \mathcal{H}_\alpha\right), \tag{12.15}$$

where $\dim \mathcal{H}_\alpha$ is the dimension of the subspace $\mathcal{H}_\alpha$. The equilibrium region is, as always, the region of maximal entropy with $\dim \mathcal{H}_{eq} \approx \dim \mathcal{H}$. In contrast to classical mechanics, however, a quantum state can be in a superposition of various macrostates, i.e., Ψ need not lie entirely in one of the subspaces making up the partition. However, we can say that Ψ realizes the macrostate corresponding to $\mathcal{H}_\alpha$ iff $\langle \Psi \mid P_\alpha \mid \Psi_0 \rangle \approx 1$, where P_α is the projection onto $\mathcal{H}_\alpha$. In Bohmian mechanics, it also makes sense to say that the macrostate of a system is determined by the branch of the wave function that actually guides the particle configuration, while disregarding empty branches.

Remark (Decoherence)**.** On the other hand, the branching of the wave function –

also known as *decoherence* – is itself an important example of a thermodynamically irreversible process. Heuristically, this can be understood as follows: Reversing the thermodynamic evolution of a Newtonian macro-system would require an exact reversal of $\sim 10^{24}$ particle velocities. Similarly, bringing two macroscopic wave packets back into interference – i.e., to overlap on configuration space – would require precise control over $\sim 10^{24}$ phases of the one-particle wave components. Simply put, each dimension of configuration space is one along which the wave packets might fail to "meet," which makes it practically impossible to bring macroscopic wave functions back into interference. Unfortunately, I am not aware of any treatment that makes the connection to the Boltzmann entropy (12.15) precise.

Returning to the arrow of time in Bohmian mechanics, it is usually assumed that the wave function of our universe is atypical, having started out in a macro-region of very low (quantum) Boltzmann entropy, while relative to this wave function, the particles are in quantum equilibrium. This can account for both the Born rule and thermodynamic irreversibility but raises the usual worries about the Past Hypothesis, referring, in this case, to the special initial macro-conditions of the quantum state (cf. Chen (2018) for a discussion and interesting solution). However, Bohmian mechanics opens up the intriguing possibility of a stationary wave function of the universe, satisfying a "constraint" Schrödinger equation of the form

$$H\Psi = 0 \tag{12.16}$$

(that would correspond to the Wheeler-de-Witt equation in canonical quantum gravity). In contrast to orthodox quantum theories, Bohmian mechanics would not be hit by the "problem of time" (see, e.g., Kiefer (2015)). The particle configuration and the effective wave functions of subsystems would, in general, evolve even if the universal wave function did not. This option seems particularly attractive if one favors a nomological interpretation of the wave function according to which Ψ is part of the physical laws rather than a *beable* over and above the particles (Dürr et al., 1997; Goldstein and Zanghì, 2013; Esfeld et al., 2014; Esfeld, 2014b). Hence, the ideal solution, from my point of view, would be something like the following: a natural (possibly non-normalizable) stationary wave function that makes a thermodynamic arrow typical (à la Carroll) with respect to some internal physical time-parameter (maybe the scale factor of spacetime?) playing the role of a universal "clock." Of course, physics cares little about my wishes, and to date, this is merely the statement of a speculative research program.

Why Determinism?

> The appearances are a sight of the unseen.
>
> — Anaxagoras, Fragment 21a

Since the empirical predictions of Bohmian mechanics are based on the Born rule, i.e., the quantum equilibrium distributions, they will always agree with the predictions of standard quantum mechanics – whenever the latter are well-defined. Now we could point to various cases in which the orthodox predictions – in contrast to the Bohmian ones – are indeed unclear or ambiguous[4], and others in which observed phenomena are much more naturally interpreted in Bohmian terms[5]. At the end of the day, however, the empirical import of the Bohmian theory comes always from its statistical analysis rather than the exact particle trajectories described by the deterministic guiding equation.

It is usually regarded as an innovation of quantum physics that the phenomena are necessarily "random" in this sense. The truth is that quantum theory entails rigorous limits on our epistemic access to the micro-cosmom (see, in particular, the Bohmian theorem of *absolute uncertainty* discussed at the end of Ch. 8.2, and Cowan and Tumulka (2016) for an analogous result about collapse theories), but the idea that according to classical mechanics (conceived as an atomistic theory) the situation was radically different as a practical matter has always struck me as naive. Why would anyone have ever thought it feasable to determine the initial conditions of Newtonian particles with infinite precision? Indeed, we have seen that even the ostensibly deterministic phenomena that are successfully predicted by Newtonian physics must ultimately be understood as typical regularities of many-particle systems. What Einstein said about Brownian motion in his 1910 lecture "Über das Boltzmann'sche Prinzip und einige unmittelbar aus demselben fliessender Folgerungen"[6] thus applies to Newtonian trajectories, in general, and translates perfectly to the situation in Bohmian mechanics:

> If we now conclude by asking once again the question "Are the observable physical facts completely causally linked with one another?", we must firmly deny it. [...] According to the theory, one would need, in order to [compute the trajectories], to know the position and velocity of every single molecule, which seems impossible. Nevertheless, the laws of mean values that have proven themselves all over, as well as the statistical laws of fluctuations applicable in those areas of subtle effects, convince us that we must adhere to the principle of a complete causal link between the occurrences in the theory, even if we cannot hope to ever obtain direct confirmation of this view through refined observations of nature. (Translation by the author.)

In other words, what gives us trust in the microscopic theory is not primarily the intuitive appeal that determinism and particle trajectories may or may not have, but the naturalness and coherence with which it grounds the empirical (statistical) phenomena. For most practical purposes, however, nothing is lost by simply postulating,

[4] E.g., arrival time measurements (Das and Dürr, 2019), or "Wigner's friend" experiments (Lazarovici and Hubert, 2019).

[5] Such as weak measurements of particle trajectories, cf., e.g., Dürr and Lazarovici (2020, Ch. 8).

[6] Archived by *Physikalische Gesellschaft Zürich.* http://www.pgz.ch/history/einstein/index.html

rather than deriving, a bunch of phenomenological and probabilistic rules – which is essentially what standard quantum mechanics is doing (Maudlin (2019) calls it the "quantum recipe").

The final irony is now that while Einstein's derivation of Brownian motion proved instrumental in fostering the acceptance of the atomic theory, mainstream physics has long since regressed to the Machian positivism that called atomic particles a "figment of the imagination" (*Hirngespinste*). No physicist today will flat out deny the existence of atoms and more elementary particles, but many will at least pay lip service to the idea that "particle" refers only to some abstract state characterized by its observable properties.

In any case, one could certainly have a general argument about whether a deterministic theory is always preferable *ceteris paribus* because it provides a more complete description of nature's course. This point, however, is somewhat mute since the situation we are facing in quantum theory is, though marked by underdetermination, not quite a *ceteris paribus* situation. The objective collapse models that are seriously entertained today make certain predictions that differ, in principle, from those of unitary quantum mechanics, while various "interpretations" of the standard formalism, which keep insisting on irreducible randomness in some form or the other, operate with a very different standard of conceptual clarity and mathematical rigor. The one alternative we have to discuss in more detail is the Everettian Many Worlds theory, which – though fundamentally deterministic – provides a description of nature that differs radically from the Bohmian one.

12.4 The Many Worlds Theory

Hugh Everett III. is credited as the father of the Many Worlds theory, although the name was only later introduced by Bryce DeWitt, and it is historically disputed whether Everett truly believed in the reality of many worlds or was, in fact, more of an instrumentalist (see Barrett (2001)). Undisputed is Everett's insistence that we must take quantum theory seriously on all scales. He thus introduced the concept of the *universal wave function* that already played a crucial role in earlier chapters (Everett, 1956). Everett recognized that the shifty split between the microscopic quantum regime and the macroscopic classical regime couldn't stand if quantum mechanics was supposed to provide a coherent description of nature. In contrast to David Bohm (1952a,b), however, Everett refused to introduce additional variables into the theory. He insisted on "pure wave mechanics" defined only in terms of the (universal) wave function and the linear Schrödinger equation. Today, it is generally accepted that such a theory results in a Many-Worlds picture, in which the decoherent branches of the wave function describe a multitude of different but equally real macro-histories.

One may certainly find such a theory bizarre or extravagant. John Bell (2004) called it "above all ... extravagantly vague" (p. 194). Indeed, on closer examination, the problem with the Many Worlds theory is not the many worlds but the question, how the theory is supposed to describe any world at all. In other words: how do we locate tables and chairs and cats and measurement devices with pointer positions in the wave function of the universe?

A first approximation to a solution goes something like this: Any point in $3N$-dimensional configuration space describes the positions of N particles in three-dimensional space, and this configuration can be such that it forms a table, or a cat, or a measurement device whose pointer points to the left. And other points in configuration space that lie nearby will describe particle configurations that deviate only slightly and will thus look macroscopically the same. In this way, we can identify entire regions of configuration space with certain macroscopic "images" that are coarse-grained from particle configurations. Thus, if a particular branch of the wave function is (suitably well) localized in a region of configuration space that coarse-grains to a cat, we can say that we have located a cat in the wave function. And if another part of the wave function is (suitably well) localized in a region of configuration space that coarse-grains to a dead cat, we can say that we have located a dead cat in the wave function.

The problem with this response doesn't even lie in the expression "suitably well" that the reader may find justifiably suspicious. The problem with this response is that we have been cheating all along. For what justifies the identification of points in $3N$-dimensional configuration space with configurations of N hypothetical particles in 3-dimensional space? What even justifies the name "configuration space" for the high-dimensional space on which the universal wave function lives? Configuration of what? If the ontology of quantum mechanics is supposed to be the wave function and the wave function alone, we cannot suddenly pretend that its degrees of freedom refer, somehow, to particle positions. As de Broglie remark already in 1927: "It seems a little paradoxical to construct a configuration space with the coordinates of points that do not exist." (Quoted from (Bacciagaluppi and Valentini, 2009, p. 346))

In the modern literature, one thus finds another strategy that falls under the philosophical concept of functionalism. The idea, in a nutshell, is that being a cat is not to be a cat-shaped configuration of matter. To be a cat is to act like a cat: to chase after a mouse when it passes by, to purr when being caressed, to land on the feet when jumping out the window, etc. The locate a cat or a table or a chair in the universal wave function is thus not to find something that *composes* a cat or a table or a chair (as Bohmian particles would), but to identify certain patterns in the wave function that, in their interplay, satisfy the pertinent functional definitions. Since quantum mechanics describes interactions first and foremost on the level of the wave function, one may expect (or hope) that the wave function will show the right dynamical behavior to represent our world (and many others like it) in this way. And the hope is somewhat substantiated by results about the classical limit of quantum mechanics which show that, in certain situations, well-localized wave packets typically propagate, to a good approximation, like classical Newtonian bodies.

There are serious doubts about whether such a wave function functionalism can work in practice, or even in principle (see, e.g., Monton (2006), Maudlin (2010)). We experience the physical world as matter moving and interacting in three-dimensional space, and it is not hard to understand how macro-objects like tables and cats could be functionally realized by microscopic objects – like particles – moving and interacting in three-dimensional space. (Indeed, Bohmians are also "macro-object functionalists" (Lewis, 2007) in this sense.) But the degrees of freedom of the wave function do not move and interact in three-dimensional space. They do not even stand in distance relations to one another but live in different dimensions of the high-dimensional "configuration" space. Hence, a cat-wave cannot chase a mouse-wave anymore than the latitude of my hand can chase its longitude. The *spacetime state realism* proposed by Wallace and Timpson (2010) (see also Wallace (2012a)), which tries to conceive the Many Worlds theory in spatiotemporal terms, might fare better in this respect. But the more elaborate the Everettian functionalism becomes, technically and metaphysically, the less plausible the claim that it is all "just unitary quantum mechanics" and thus somehow more economical than Bohmian mechanics or collapse theories.

Probabilities in the Many Worlds theory

The Many Worlds theory has trouble reproducing the statistical predictions of quantum mechanics, i.e., the Born rule. The problem is not that the theory is deterministic (there is only one equation, the Schrödinger equation, which is deterministic). We have discussed in detail how objective probabilities can be grounded in deterministic laws, including the derivation of Born's rule in Bohmian mechanics. The critical question when it comes to probabilities in the Many Worlds theory is rather: probabilities of what? The theory says, after all, that *every* possible result of a quantum experiment actually obtains.

If we consider, for example, a spin-measurement on a spin-1/2-particle, it doesn't seem meaningful to ask for the probability of measuring "Spin UP" or "Spin DOWN." In one world, the upper detector clicks and we measure spin UP, in another world, the lower detector clicks and we measure spin DOWN (assuming, of course, that the particle was not in an eigenstate).

Naively, one may think that quantum statistics refer to the relative frequency of worlds. One outcome being "more likely" than another simply means that it will be realized in a greater number of world-branches. However, if this were true, the Many Worlds theory would make incorrect predictions. Suppose our electron is in the spin-state

$$\psi_1 = \frac{1}{2}|\uparrow_z\rangle + \frac{1}{2}|\downarrow_z\rangle$$

and we measure its spin in z-direction. At the end of the experiment, our world will have split into two branches: one in which we have measured "spin UP," and one in which we have measured "spin DOWN." Each possible outcome thus occurs in an equal number of worlds, in accordance with the quantum mechanical probabilities. But now suppose the electron is instead in the spin-state

$$\psi_2 = \frac{1}{\sqrt{3}}|{\uparrow_z}\rangle + \sqrt{\frac{2}{3}}|{\downarrow_z}\rangle.$$

According to Born's rule, the probabilities are $\frac{1}{3}$ for "spin UP" and $\frac{2}{3}$ for "spin DOWN." According to the Many Worlds theory, however, the outcome is the same as before: we end up with two branches, in which the result of the measurement is "spin UP" and "spin DOWN," respectively. Hence, naive "branch counting" – to the extent that it even makes sense – doesn't yield statistics consistent with quantum mechanical predictions. Notably, the "weights" of the world-branches, i.e., the pre-factors $c_1 = \frac{1}{\sqrt{3}}$ and $c_2 = \sqrt{\frac{2}{3}}$, have no physical significance. It's not like one world is "more real" or "exists with higher intensity" than the other (*pace* Vaidman (2018) who seems to argue in this way); the functional relations within a branch are all that matters.

In light of our previous discussions, the reader may already realize that the difficulties come, at least in part, from the idea that probabilities pertain to individual events like the outcomes of a particular measurement. In any case, since many authors find it difficult to identify interesting probabilities in an Everrettian multiverse, they try to locate them in our head, that is, to interpret them as *subjective* probabilities. For instance: after I perform a spin measurement – but before I look at the detector to see the result – I don't know if I find myself in a world in which the detector has registered "spin UP" or a world in which it has registered "spin DOWN." What should be my "degree of belief" for one or the other? If someone offers me a 2:1 bet on "spin UP," should I take it? The "chances" in this case arise from my "self-locating uncertainty" (Sebens and Carroll, 2016): I do not know which branch of the Many Worlds universe my present self inhabits and the goal of a theoretical analysis would be to show that it is rational to assign degrees of belief according to the quantum mechanical probabilities. Other authors have taken a more decision-theoretic perspective, trying to argue that it is rational, in an Everettian multiverse, to act according to the Born probabilities. In this vein, Wallace (2012b) stipulates a set of 10 axioms to justify the use of the branch amplitudes for calculating expected utilities in decision problems. Maudlin (2014) points out that these axioms do not allow a rational agent to split a payoff among two or more of her future copies, i.e., to see any utility in the option *all of the above*. "If one were mischievous, one might even put it this way: Wallace's 'rationality axioms' entail that one should behave as if one believes that Everettian quantum theory is false." (p. 804)

But regardless of how convincing such rationality principles may or may not be, there is something unsatisfactory about retreating to dutch-book arguments or purely subjective probabilities. After all, in our laboratories, we do not take bets or poll scientists on their personal expectations. We observe concrete statistical regularities that can be reproduced in independent experiments and are very well predicted by

Born's rule. A quantum theory should be able to account for these empirical facts – otherwise, the theory is no good.

Everett's typicality argument

Hugh Everett's own explanation of the Born rule (which, oddly, has been abandoned by modern Everettians) was based on a typicality argument – and thus on objective probability assignments – not unlike to the one we have discussed in Bohmian mechanics. Therein, the $|\Psi|^2$-measure determined by the universal wave function (that is, the absolute squares of the pre-factors of the various branches) defines a typicality measure on world-branches, which is then used to identify statistical regularities that hold in the overwhelming majority of branches.

The typicality measure is thereby distinguished by something akin to stationarity under the Schrödinger evolution. More precisely, Everett stipulated three requirements for the typicality measure (Barrett, 2016):

1. It should be a positive function of the complex-valued coefficients associated with the branches of the superposed wave function.

2. It should be a function of the amplitudes of the coefficients alone.

3. It should satisfy on additivity requirement: if a branch b is decomposed into a collection $\{b_i\}$ of sub-branches, the measure assigned to b should be the sum of the measures assigned to the sub-branches b_i.

This last additivity condition can be understood diachronically as *stationarity*, in the sense that the weight associated with any world at any time equals the sum of weights associated with all of its branching histories at later times. As Hugh Everett explained:

> We wish to make quantitative statements about the relative frequencies of the different possible results of observation – which are recorded in the memory – for a typical observer state; but to accomplish this we must have a method for selecting a typical element from a superposition of orthogonal states. [...] The situation here is fully analogous to that of classical statistical mechanics, where one puts a measure on trajectories of systems in the phase space by placing a measure on the phase space itself, and then making assertions ... which hold for "almost all" trajectories. [...] However, for us a trajectory is constantly branching (transforming from state to superposition) with each successive measurement. To have a requirement analogous to the "conservation of probability" in the classical case, we demand that the measure assigned to a trajectory at one time shall equal the sum of the measures of its separate branches at a later time. This is precisely the additivity requirement which we imposed and which leads uniquely to the choice of square-amplitude measure. (Everett, 1957, pp. 460-461)

Consider for instance a sequence of z-spin-measurements performed on identically prepared electrons in the state

$$\alpha|{\uparrow_z}\rangle + \beta|{\downarrow_z}\rangle, \quad |\alpha|^2 + |\beta|^2 = 1.$$

We denote by $|{\Uparrow}\rangle$ respectively $|{\Downarrow}\rangle$ the state of the measurement device (and, in the last instance, the rest of the world), that has registered "spin UP," respectively "Spin DOWN." After the first measurement, our world splits according to

$$\alpha|{\Uparrow}\rangle|{\uparrow_z}\rangle_1 + \beta|{\Downarrow}\rangle|{\downarrow_z}\rangle_1\,, \tag{12.17}$$

where the index 1 indicates the measurement on the first particle. With the second measurement, each world splits anew, namely according to the decoherent wave branches:

$$\alpha^2|{\Uparrow\Uparrow}\rangle|{\uparrow_z}\rangle_2|{\uparrow_z}\rangle_1 + \beta\alpha|{\Downarrow\Uparrow}\rangle|{\downarrow_z}\rangle_2|{\uparrow_z}\rangle_1 + \alpha\beta|{\Uparrow\Downarrow}\rangle|{\uparrow_z}\rangle_2|{\downarrow_z}\rangle_1 + \beta^2|{\Downarrow\Downarrow}\rangle|{\downarrow_z}\rangle_2|{\downarrow_z}\rangle_1$$

The first three steps of the branching process are shown in the following figure 12.2:

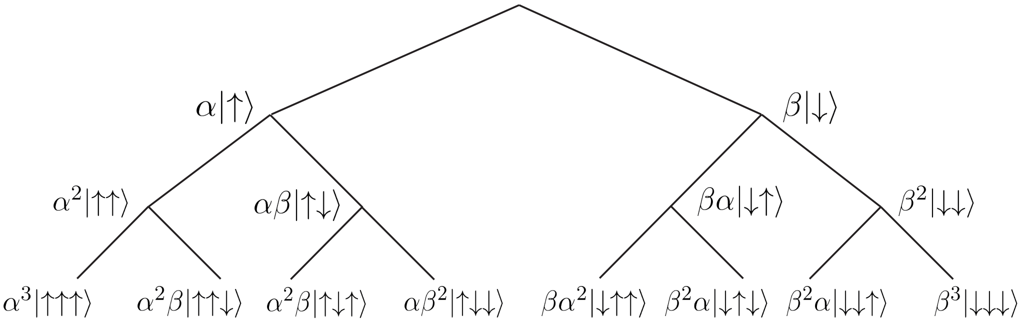
$\alpha|{\uparrow}\rangle$
$\beta|{\downarrow}\rangle$
$\alpha^2|{\uparrow\uparrow}\rangle$
$\alpha\beta|{\uparrow\downarrow}\rangle$
$\beta\alpha|{\downarrow\uparrow}\rangle$
$\beta^2|{\downarrow\downarrow}\rangle$
$\alpha^3|{\uparrow\uparrow\uparrow}\rangle$
$\alpha^2\beta|{\uparrow\uparrow\downarrow}\rangle$
$\alpha^2\beta|{\uparrow\downarrow\uparrow}\rangle$
$\alpha\beta^2|{\uparrow\downarrow\downarrow}\rangle$
$\beta\alpha^2|{\downarrow\uparrow\uparrow}\rangle$
$\beta^2\alpha|{\downarrow\uparrow\downarrow}\rangle$
$\beta^2\alpha|{\downarrow\downarrow\uparrow}\rangle$
$\beta^3|{\downarrow\downarrow\downarrow}\rangle$

Figure 12.2: Branching Many Worlds histories after three spin measurements. Successive arrows indicate successive outcomes. (Graphic adapted from Barrett (2016))

The conservation of the measure in each branch can now be readily verified. For instance, along the history on the left, we have after the second measurement:

$$|\alpha|^4 + |\alpha|^2|\beta|^2 = |\alpha|^2(|\alpha|^2 + |\beta|^2) = |\alpha|^2.$$

This conservation of the typicality measure wouldn't hold if we weighted each branch equally, i.e., performed a simple branch counting. This is easy to see if we assume that, in our example, the second measurements in the already separated worlds occur at different times. If the second measurement occurs earlier in the left branch than in the right branch, the total number of worlds first increases from 2 to 3, and the weight of the right branch suddenly drops from 1/2 to 1/3. That is, until the second measurement occurs in the right branch as well, resulting in a total of 4 branches.

Following the principle of stationarity, we thus arrive at the $|\Psi|^2$-measure as the typicality measure by which we weight branches of the universal wave function corresponding to Everettian worlds. And according to this measure, a *typical* branch will be one in which Born's rule – and thus quantum statistics – holds.

Let's check this for our go-to example of consecutive spin measurements. After n measurements, the measure of worlds in which the outcome "spin UP" occurred exactly k-times is $\binom{n}{k}|\alpha|^{2k}|\beta|^{2(n-k)}$. Writing $|\alpha|^2 =: p$ and $|\beta|^2 = 1 - p$, we see that this is a Bernoulli process with n independent trials and "success" probability p. According to the law of large numbers, the *typical* relative frequencies for *spin UP* are thus $\frac{k}{n} \approx p = |\alpha|^2$, matching the prediction of quantum mechanics.

In the upshot, Everett's analysis establishes that quantum statistics hold across *typical histories* of the constantly branching Many Worlds universe. We would now like to conclude this analysis with a solid probabilistic prediction and say something like: "So, *I* should expect to experience a typical history in which the Born statistics hold." However, to whom this *I* refers in a Many Worlds universe is a very subtle and difficult question. Even understood in a branch-indexical way, it does not pick out a unique future macro-history. My current branch is going to split repeatedly, and there will be a great many (maybe infinitely or unquantifiably many) versions of *me* registering different outcome statistics.

Hence, I see no way around the conclusion that the Many Worlds theory lacks a *predicitve* quality in this sense. I do believe, however, that Everett's typicality reasoning can ground post-factum *explanations*. When I lie on my death bed and wonder, with my last breath, why I have experienced a history consistent with quantum statistics, I would die in peace knowing that this is typical; that *nearly all* versions of me existing in the Everettian multiverse have experienced a history consistent with quantum statistics.

Wilhelm (2019) makes the interesting observation that this typicality explanation is manifestly distinct from probabilistic explanations if we agree that the latter presuppose that only one of the possible outcomes is actually realized:

> "[I]n Everettian quantum mechanics, the various possible outcomes of any given experiment all obtain. Everett himself makes this point: it would be a mistake, he says, to think of just one outcome as obtaining, to the exclusion of the others. So the sequences of outcomes other than the one invoked in the explanandum ... occur too. But in probabilistic explanations, that cannot happen. In probabilistic explanations, the event invoked in the explanandum is the only outcome, of the various possible mutually exclusive outcomes, that occurs."

One could try to evade Wilhelm's argument by falling back on self-locating probabilities: only one of the copies of D.L. existing in the multiverse is the branch-indexical *I*.

But me being me doesn't seem like the right explanandum. There is no self-location uncertainty in the death bed scenario; I know what life I have lived, i.e., what branch I am actually on. For better or worse, the typicality explanation ends with the fact that the Born rule holds across the great majority of branches. To ask further, for the probability that I find myself on any one of the branches (as if my *ego* had been somehow thrown at random into the multiverse) strikes me as redundant at best and meaningless at worst.

Part III

Metaphysics

Chapter 13

Typicality and the Metaphysics of Laws

> It would seem unreasonable ... if the whole universe and each and every part of it were in order..., while there were nothing of the kind in the principles.
>
> — Theophrastus, *Metaphysics* 7 a 10[1]

13.1 What are Laws of Nature?

Over the past few decades, the best system account has developed into a popular, maybe even the dominant, position regarding the metaphysics of laws of nature. In brief, this view holds that laws of nature are merely descriptive, an efficient summary of contingent regularities that we find in the world. Metaphysically, it is based on the thesis of Humean supervenience – named in honor of David Hume's denial of necessary connections – that David Lewis (1986a) famously characterized as "the doctrine that all there is to the world is a vast mosaic of local matters of particular fact, just one little thing and then another." Laws of nature are then supposed to supervene on this Humean mosaic as the deductive system that strikes the optimal balance between simplicity and informativeness in describing the world.

The Humean "regularity view" of laws is opposed to the "governing view," in its various forms, according to which the fundamental laws play an active role in guiding, or producing, or constraining the history of the universe. For the purpose of our discussion (and avoiding a complete overview of the various anti-Humean positions), I will take the main contemporary contenders to be dispositional essentialism (Bird, 2007) – which grounds the laws of nature in dispositional properties instantiated by the fundamental entities – and nomic primitivism (Maudlin, 2007a), which admits "law of nature" as a primitive ontological category, and laws as fundamental entities into the ontology of the world.

Notably, our discussion will only be concerned with fundamental physical laws that

[1] Quoted after Finkelberg (2017, p. 59).

govern or summarize the entire physical history of the world, although both Humean and anti-Humean views may be compatible with more deflationary notions.[2] On the other hand, I am only going to defend the minimal anti-Humean thesis that laws "govern" or "constrain" that history. If the view that laws "produce" entails more than that (as Schaffer (2016) argues) or is tied to a particular metaphysics of time (see Loewer (2012b) versus Maudlin (2007a)), it will require independent justification.

There is one way to phrase the debate between Humeanism and anti-Humean metaphysics that I find both uninteresting and misleading (for reasons that will become clearer in the course of our discussion): Laws can determine regularities (as their instances), and regularities can determine laws (as their best systematization), and so the question is: what comes first, what is more fundamental, the regularities or the laws? ("What grounds what?" is how one would put it, more properly, in contemporary metaphysics, see, e.g., Schaffer (2008, 2009).) One could then be skeptical about one of the two grounding relations and choose sides on this basis; e.g., deny that there can ever be an unambiguously best system, or find it utterly mysterious how laws are supposed to "govern" anything. This is not my main concern, however, and our discussion will grant that both the regularity and the governing view of laws are at least conceptually sound. Instead, I consider the debate between Humean and anti-Humean metaphysics to be first and foremost a debate about fundamental ontology – whether there is more to the fabric of the world than the Humean mosaic – and the interesting choice to be one between ontological parsimony and other theoretical virtues.

In this debate, Humeans have had remarkable success in defending a prima facie implausible position against all objections that have been thrown their way (and then claim victory on the grounds of parsimony). In recent years, criticisms of the best system account have focussed, in particular, on the lack of explanatory power of Humean laws (e.g., Maudlin (2007a); Lange (2013)), the alleged subjectivity of the best system (Armstrong, 1983; Carroll, 1994), or the commitment to a separable ontology which is put into question by the entanglement structure of quantum mechanics (Maudlin, 2007a). Humeans have resisted all of these attacks with some, though varying, degree of persuasiveness (see, e.g., Lewis (1994); Loewer (1996, 2012b); Cohen and Callender (2009); Hall (2009), for the application of Humeanism to (Bohmian) quantum mechanics see Esfeld et al. (2014); Esfeld (2014b); Miller (2014); Callender (2015); Bhogal and Perry (2017)). This is not to say that the objections have no merit, but I believe they have not quite managed to capture the implausibility of Humean metaphysics and turn it into a compelling argument for modal realism.

The present chapter aims to do just that. It will thereby elaborate on a fairly common anti-Humean intuition, which is to look at the astonishing order in our cosmos, the uniformity of nature expressed by the simple and successful laws discovered in physics, and ask: *how likely is it that these regularities come about by chance?*

[2] Maudlin's view, according to which laws produce the history of the universe, is arguably not.

One place where this argument is articulated in some detail is in the book "The Divine Lawmaker" by John Foster (2004):

> What is so surprising about the situation envisaged – the situation in which things have been gravitationally regular for no reason – is that there is a certain select group of types, such that (i) these types collectively make up only a tiny portion of the range of possibilities, so that there is only a very low prior epistemic probability of things conforming to one of these types when outcomes are left to chance ... (p. 68)

I agree with the basic intuition but believe that the argument, thus phrased, cannot succeed. Indeed, Humeans have several good points to make in response:

1. We do not have to account for why the law of gravitation – or any other particular law described by physics – holds in our universe. Anti-Humean views do not explain this, either. The debate is about what it is to be a law, not why the laws of our world are what they are.

2. What do you mean by "chance"? The thesis of Humean supervenience holds that the history of the universe, the distribution of "local particular facts," is contingent. But contingency, or the absence of a further metaphysical ground, does not mean randomness. In fact, Humean metaphysics are opposed to all intuitions about the mosaic being "produced" by a chancy process – particles performing random motions, or God playing blindfolded darts and throwing local particulars into spacetime, or anything like that.

3. Where do your "prior probabilities" come from? What determines the right probability measure over possible worlds? All successful applications of probability theory come from *within* science. And according to the most prominent Humean account (see Ch. 5), the fundamental probability measure that grounds probabilistic predictions and rational priors is itself part of the best system that supervenes on the Humean mosaic. In other words: the actual world determines all relevant probabilities; there are no justified a priori probabilities which could warrant the conclusion that a world like ours is unlikely.

These points are well taken. In particular, I agree that references to probability or chance are dubious in a metaphysical context where subjectivist, frequentist, and regularity interpretations all seem questionable or inappropriate. The concept of *typicality*, however, strikes me as a perfect fit for the issue at hand.

13.2 Typicality in Metaphysics

Why does typicality avoid the objections raised against probability? For one, because typicality statements are extremely robust against variations of the measure used to quantify subsets of a reference set W, so much so that, in many cases, the question how to pick the right measure or what it even means to be the "right" measure doesn't even arise (cf. Maudlin (2007b, p. 286)).

Our very first example of a typicality statement was: *almost all real numbers are irrational.* That is, being irrational is typical within the set $\mathbb{R}$ of reals.

In what sense is this true? First and foremost, in terms of cardinalities. The set of real numbers is uncountably infinite, while the subset of rational numbers is only countably infinite. Therefore, $\frac{|\mathbb{R}\setminus\mathbb{Q}|}{|\mathbb{R}|} = 1$ and $\frac{|\mathbb{Q}|}{|\mathbb{R}|} = 0$. This is a very precise and generally uncontroversial sense in which almost all real numbers are irrational. In principle, nothing more needs to be said here. However, since we will use it later on – and since its more familiar from applications in physics – we can spell out typicality in terms of a measure in the sense of mathematical measure theory. It then seems natural to consider the uniform Lebesgue measure on $\mathbb{R}$, which makes it true that *all real numbers except for a subset of measure zero* are irrational. (This is weaker than the statement in terms of cardinalities; there exist Lebesgue null-sets that are uncountably infinite.) Note that the Lebesgue measure on $\mathbb{R}$ is not normalizable, so it cannot be confused with a probability measure. But maybe the uniform measure is suspicious as it reeks too much of a "principle of indifference." Fair enough, we can pick virtually any other measure we like. Any non-discrete measure, i.e., any measure that is zero on singletons, will agree that $\mathbb{Q} \subset \mathbb{R}$ has measure zero. (By σ-additivity, a measure can be non-zero on countable sets if and only if it is non-zero on some one-element sets.) Simply put, we assume nothing more than that a one-element subset is vanishingly small compared to an uncountably infinite set. There is thus a very innocent and intuitive sense in which all reasonable measures agree on the meaning of "typical."

Typicality statements in physics usually admit exception sets of very small (but non-zero) measure. Here and in our following discussion, we can use a particularly strict standard of typicality that provides even stronger results than can be realistically obtained in the physical context.

Another crucial difference between typicality and probability is that typicality is not tied to ignorance, randomness, or indeterminism. It is an objective, determinate fact that typical real numbers are irrational. It has nothing to do with anyone's credences, nor with some number being picked at random, or picked out at all.

When applied to a reference class of possible worlds, typicality figures in a way of reasoning about contingency. (And contingency, if anything, is central to Humean metaphysics.) We have already applied this in the context of physics and statistical mechanics; now I am going to argue that typicality extends to a powerful way of reasoning in metaphysics.

The typicality fact that the best system account has to deal with is then the following: It is typical for Humean worlds to have no Humean laws. Almost all Humean worlds *do not have any regularities in the first place* but are too complex to allow for a systematization by physical laws. (This will be rigorously proven for deterministic laws and in a more hand-waving fashion for probabilistic ones.) The challenge to Humean metaphysics is thus not to account for why we find *these particular* laws in our universe but why we find any laws at all. Conversely, if we do live in a world that is regular enough to be described by laws of nature, the best explanation is the existence of something in the fundamental ontology that makes it so.

Ontological possibility

> The orthographical symbols are twenty-five in number. This finding made it possible, three hundred years ago, to formulate a general theory of the Library and solve satisfactorily the problem which no conjecture had deciphered: the formless and chaotic nature of almost all the books.
>
> — Jorge Luis Borges, The Library of Babel

While typicality statements, at least when made with respect to an infinite reference set, require some mathematical tool like measures to give precise meaning to the locution "almost all," their truth-maker – and what is ultimately doing the explanatory or argumentative work – is not the measure but the reference class with respect to which properties come out as typical or atypical (or neither). In particular, an important (if not the most important) way in which laws of nature, however conceived, *explain* or *predict* is by delimiting a set of nomologically possible worlds that makes certain physical phenomena typical.

Indeed, we learned with the breakthrough of atomism and the development of statistical mechanics that, due to the huge number of microscopic degrees of freedom, the fundamental dynamical laws allow for vast possibilities far beyond what had been thought of as permissible by physical laws. That apples fall to the ground but don't spontaneously jump up, that planets do not suddenly fly off their orbit (while emitting an ultra-fast particle in the opposite direction), and heaps of dust do not spontaneously transform into dinosaurs, is explained not by the fact that such events are nomologically impossible but by the fact that they are atypical, i.e., they would require extremely special micro-conditions of the universe.

Analogously, if we want to apply a typicality reasoning in a metaphysical context – evaluating the merits of a Humean versus anti-Humean metaphysics – we need a reference class of possible worlds that is determined by the respective ontologies and does not a priori coincide with nomic possibilities. The relevant reference class that I propose is generated as follows:

Fix the fundamental ontology of the world as postulated by a metaphysical theory, that is, the fundamental entities with their essential properties, and consider all their possible configurations, i.e., possible distributions of contingent properties (such as spatiotemporal relations) over these "individuals."

Possible worlds thus generated are sometimes called Wittgenstein worlds, in reference to the following passage of the *Tractatus*:

2.0271 The object is the fixed, the existent; the configuration is the changing, the variable.

2.0272 The configuration of the objects forms the atomic fact. [...]

2.04 The totality of existent atomic facts is the world.

Allowing for "augmentation" and "contraction" – adding individuals (but not universals) beyond those that exist, or removing some that do – the set of Wittgenstein worlds is extended to "Armstrong worlds" (Kim, 1986) and the theory of modality known as Combinatorialism (Armstrong (1986, 1989); see Sider (2005) for a recent discussion). In our discussion, we will not need augmentations and contractions, and if we consider the option that laws of nature may themselves be among the fundamental "entities" that exist in our world, adding or removing them would defeat the purpose. Hence, we shall keep the basic furniture of our world fixed, both in type and in number. Notably, I am not interested in defending this or any other version of Combinatorialism as a full-blown theory of metaphysical possibility. Instead, let us call the relevant notion of modality *ontological possibility*, the crucial point being that a world is ontologically possible (according to a metaphysical theory) if it has the same fundamental ontology as postulated for the actual world.

Here are some examples for the use of ontological possibility: If the fundamental ontology of the world consists in point particles moving in space, it is ontologically necessary for all material objects to be spatially localized. If the fundamental ontology of the world consists of N permanent point particles, it is ontologically impossible for any object to be composed of more than N parts. According to a Super-Humean theory of space or spacetime (Huggett, 2006), it is ontologically possible for spacetime to have more than four dimensions. According to a functionalist theory of the mind – but not according to theories that postulate "minds" as ontological primitives – consciousness is ontologically contingent.

Why should we care about ontological possibility given that it is, as far as I can tell, a notion that we have just stipulated rather than an established philosophical concept?

Most basically, because this seems like a fairly standard semantic interpretation of what a hypothesis about the fundamental ontology of the world means.

Intuitively, because the fundamental entities that we believe to exist should have a distinguished epistemic and explanatory role over those that are merely possible or conceivable.

Most importantly, because ontological possibility, thus defined, is the form of modality that captures the disagreement between Humean and many anti-Humean metaphysics. Humeans and anti-Humeans will agree on the set of nomological possible worlds (if they agree on what the best theories of physics are) and they may agree or disagree on metaphysical possibility for all kinds of philosophical reasons that can go beyond their stance on laws of nature. Humeans, however, are committed to a *principle* *of unrestricted recombination* (Lewis, 1986b): it is possible to change the configuration of fundamental entities or properties in any part of the Humean mosaic while holding fixed the rest of the mosaic. This is the positive content of Humean metaphysics, the flip side of the negative theology regarding necessary connections and all kinds of "non-Humean whatnots."

The main anti-Humean positions, on the other hand, hold that there exists something in the actual world – be it essential dispositional properties or primitive laws – that restricts combinations; that makes it impossible, let's say, for a world to have the same fundamental ontology as ours but a distribution of masses incompatible with the law of gravitational attraction. Notably, the relevant ontological commitment is not to *some* non-Humean laws, no matter how silly or complex, but to the fundamental laws that physics discovers in our universe (and of which, as of today, we have only partial or approximate knowledge). The anti-Humean positions I have in mind also include the view that the manifestation of the primitive laws or dispositions are essential to them, i.e., that a non-Humean law is the same in every world in which it exists. (This may not extend to certain parts of the laws, like the constants of nature figuring in their formulation. Typicality considerations then give rise to the issue of fine-tuning of the constants, which is indeed a big topic in fundamental physics but beyond the scope of our discussion.) The different meanings of "nomological possibility" under a Humean and anti-Humean understanding is then manifested in the fact that according to the latter but not the former, ontologically possible worlds are a subset of the nomologically possible ones. Of course, many anti-Humean theories go as far as claiming that nomic possibility coincides with metaphysical possibility, but this is an unnecessarily strong assumption for our purposes.

Typicality and the Case against Humeanism

With such a reference class of ontologically possible worlds, typicality can play a similar role in metaphysics as it does in the physical sciences. Any law-hypothesis in physics designates a set of nomologically possible worlds. This set must contain the actual world for the proposed law to have any chance of being true. However, this is not sufficient for us to judge the law-hypothesis as compelling or explanatory or even empirically adequate. For instance, there are very plausibly Newtonian universes which are such that whenever particles are shot through a double slit and recorded on a screen, they form an interference pattern. These and other quantum phenomena are not made impossible by Newtonian laws; they just come out as atypical. On the other hand, whenever we succeed in explaining (macroscopic) phenomena based on the fundamental (microscopic) laws, we show that they are typical, i.e., obtain in nearly all nomologically possible worlds. Among the typical regularities of our world are also statistical regularities, which is where objective probabilities come into play (see Chs. 4 and 5). And if Bohmian mechanics is true, we even understand why quantum statistics are typical in this sense (Ch. 8).

The case of the *thermodynamic arrow* discussed in Chapters 9 and 11 was a particularly interesting example. It is argued, based on the insights of Boltzmann's statistical mechanics, that nearly all possible micro-histories, relative to a low-entropy initial macrostate, correspond to an evolution of increasing entropy. However, it is atypical for the universe to be in a low-entropy state, to begin with. This is why it's generally considered necessary to invoke the Past Hypothesis as an additional theoretical postulate, and there is a big debate about whether this Past Hypothesis is of the right kind to be a basic postulate in a physical theory – even an additional law of nature – or whether it cries out for further explanation.

In general, the way in which we evaluate physical theories is thus roughly the following: we consider the set of nomologically possible worlds determined by the laws it postulates, and require that the saliant and relevant features of our world – the phenomena which are the target of explanation – come out as typical (or, at the very least, not as atypical). If our world corresponds, in the relevant respects, to an extremely special and fine-tuned, i.e., atypical, model of the theory, we amend or reject the theory. If we did not follow this standard, we would lose all means to test a theory against empirical evidence because special initial conditions could account for almost anything.

I submit that a similar standard should apply in metaphysics when we judge proposals for a fundamental ontology of the world. If we want to know what explanatory work a "metaphysical theory" is doing, and how it matches the world we live in, we should consider the set of ontologically possible worlds determined by the fundamental ontology it postulates and require, at the very least, that the features of our world that fall under the purview of the proposed metaphysics do not come out as atypical.

We will never get around the basic problem of underdetermination, of course, but this does not mean that there are no rational standards by which ontological commitments can be judged against the manifest appearance of the world. Typicality provides such a standard; and if we reject it, we could postulate virtually any ontology we like – as long as it gives us enough "degrees of freedom" to play around with – and claim that they are arranged in precisely such a way as to ground, or realize, or serve as the supervenience base of whatever structure we identify in nature. In other words: except for being logically inconsistent, both physical and metaphysical theories cannot, in general, do any worse in their respective domain than make the relevant features of our world atypical. When we consider proposals for the metaphysics of laws, the "lawfulness" of the world is undoubtedly a relevant feature.

Typicality, we recall, is associated with the following rationality principle: Suppose we accept a theory T and we come to believe that our world has a salient and relevant property P. If it turns out that P is typical according to our theory, there is nothing left to explain. If, however, it turns out that P is atypical according to T, we have to look for additional theoretical principles that provide further explanation for P, or else, in the last resort, reject our theory.

Atypicality, in other words, creates an epistemically unstable situation, and refusal to move means, in effect, to give up on a rational understanding of the world. The idea that our world just happens to be, in some relevant respect, an atypical model of our theory is unacceptable in science. I believe that this rationality principle is so deeply rooted in scientific thought that it is rarely made explicit, let alone questioned. As a matter of fact, more authors (see, e.g., Putnam (1969)) have questioned the laws of logic than entertained "explanations" based on atypicality.

Very much related to this is another precedent from science, namely that of typicality as a necessary condition for a successful reduction. For instance, we accept the reduction of the thermodynamic theory of gases to the kinetic theory of particles – including the ontological reduction of gases to particle configurations – because the atomistic theory makes the relevant gas properties typical. Conversely, since special micro-configurations could realize almost anything, the typicality standard prevents trivial and spurious accounts. Explanations of the form: Assume the initial conditions of the universe (or some relevant subsystem) are such that P, then P. Or reductions of the form: Assume that X has the right configuration to realize/ground Y, then we can reduce Y to X. Humean supervenience strikes me as having essentially this character: Assuming that the Humean mosaic is exactly as if governed by laws, we can reduce the laws to the mosaic.

It is admittedly difficult to find genuine examples of typicality reasoning in metaphysics that do not rely on natural laws and hence nomic possibilities. However, it is not too much of a stretch to revisit Leibniz' monadology, which denied the possibility of causal interactions between different monads (fundamental substances that can have either mental or material attributes) and ask: Why did Leibniz have to invoke his infamous doctrine of *pre-established harmony* to account for the coordination between physical and mental states (a form of which David Hume endorsed, as well)? Why could he not have left every monad to itself and claim that it is a contingent fact of our world that they happen to evolve in conformity? Well, because this claim is absurd, because without God's synchronization and in the absence of any causal or metaphysical connection, the conformity of mental and physical events would not be merely unexplained but atypical. Clearly, there would be countless more ways in which the mental and physical history of the world (and each person individually) could be in discord than in harmony, and clearly, discord is thus what Leibnizian metaphysics without pre-established harmony would imply. (Notably, we can make this judgment with high confidence even though we have nothing like a "probability measure" over possible mental states.) In the upshot: because his ontology of monads makes the conformity of mental and physical events atypical (though not impossible), and because giving up on this conformity would lead to absurdity or de facto solipsism, Leibniz had to postulate an additional metaphysical principle, viz. pre-established harmony.

A Humean ontology, as we shall now prove, makes the lawfulness of the world atypical, the harmony (so to speak) between physical events at different times at places that would allow for a systematization of the mosaic. As a consequence, we can accept a Humean ontology and be anti-realists about laws (which is not *completely* absurd; maybe Nancy Cartwright (1983) – though no Humean – is right, and we never had good reasons to believe that laws of nature should be exactly and universally true). Or we can believe in true universal laws and look for additional metaphysical principles that account for their existence. What most advocates of the best system account maintain, however, is that Humean metaphysics are true and, at the same time, that our world is an atypical instantiation of a Humean ontology – not just with respect to some minor detail but with respect to its lawfulness, the very feature at the center of their account. And this thesis, as a matter of reason and scientific practice, cannot be accepted.

13.3 Typical Humean Worlds have no Laws

In this section, we will prove the main theorem of this chapter. In brief: typical Humean worlds have no laws. We begin with a simple toy model that we call the *Chaitin model*, after Gregory Chaitin (2007), who, based on ideas that strike me as very Humean, proposed a connection between scientific practice and algorithmic information theory.

The Chaitin model

In our model, a world – with the totality of physical facts – is represented by an infinite sequence of 0's and 1's. Assuming a principle of unrestricted recombinations, the set of ontologically possible Humean worlds thus corresponds to $W = \{0,1\}^{\mathbb{N}}$, the set of all possible sequences.

The *Kolmogorov complexity* of a sequence $w \in W$ is defined as the length of the shortest algorithm that generates it. If w has finite Kolmogorov complexity, i.e., can be produced by a finite algorithm, it is called *algorithmically compressible.*

For instance, the sequence $w_0 = 0101010101...$ can be generated by an algorithm (an infinite loop) like

```
while True:
    print("01")
```

so that it is algorithmically compressible with Kolmogorov complexity of 22 or less.

We can think of an algorithm as a candidate for a best system law, its role being to provide an efficient summary of the world (sequence). In the spirit of the best system account, the length of the algorithm can be thought of as the measure of its simplicity. However, our argument will not require laws to be particularly simple, they only have to be finite.

One problem, also familiar from the best system account, is that the length of an algorithm depends on the language in which it is written.[3] We will call two languages

[3] The short example above is written in *Python*.

L_1 and L_2 *intertranslatable* if there exists a finite set of rules translating any algorithm in L_1 into an algorithm in L_2 and vice versa. It is easy to check that intertranslatability is an equivalence relation, and that the Kolmogorov complexity of a sequence with respect to any two intertranslatable languages differs at most by a finite constant. Hence, algorithmic compressibility is well-defined on these equivalence classes.

It is well known that the best system account would be trivial without some restriction on the admissible languages in which the systematizations can be formulated. For otherwise, the best system would simply consist in a primitive predicate F such that $F(w)$ is true if and only if w is the actual world @, see Lewis (1983, p. 367). "Being intertranslatable with any language known to humanity" seems like a very generous restriction, more so than if we assumed a privileged language of perfectly natural predicates. (In the toy-model, we could also define compressibility with respect to a universal Turing machine, but it may be less obvious how this generalizes.)

Hence, let $\mathcal{L}$ be the set of finite algorithms ("possible laws") in any language intertranslatable with some language known to humanity, and $W^* \subset W$ the corresponding set of compressible sequences. We call any $w \in W^*$ a *lawful world.*

Now, the following are simple mathematical facts:

- The set W of "possible Humean worlds" is uncountably infinite (its cardinality is that of the continuum): $|W| = 2^{\aleph_0} > \aleph_0$.
- The set $\mathcal{L}$ is countably infinite: $|\mathcal{L}| = \aleph_0$. [There are at most countably many admissible languages and countably many finite algorithms that can be formulated in each language. A countable union of countable sets is countable.]
- The set of compressible sequences ("lawful worlds") cannot be greater than the set of possible algorithms ("laws"): $|W^*| \leq |\mathcal{L}| = \aleph_0$. [Since each algorithm generates at most one sequence.]
- We conclude: $\frac{|W^*|}{|W|} = 0$. Hence, almost all sequences are algorithmically incompressible. Or: Almost all Humean world have no laws.

As in the example of irrational numbers, we could also express typicality in terms of a measure rather than cardinalities. It then holds true that $\mu(W^* \subset W) = 0$ with respect to *all* measures on W that are zero on singletons. In the upshot, "lawfulness" is atypical among Humean worlds under any reasonable interpretation of the concept.

From the toy model to the real world

While I hope the model to be instructive, the real world is evidently not a sequence of numbers, and fundamental laws of nature are not just algorithms for data compression but, first and foremost, dynamical laws for the microscopic constituents of the world. In order to extend our previous result to realistic physical laws – focusing, for now, on deterministic ones – we proceed as follows:

We fix a slice V of the mosaic which is sufficiently extended in space and time as to fix not just initial conditions for any deterministic dynamics, but also the values of all free parameters, like constants of nature, that may appear in their formulation. (V could be the actual history of our universe up to some time t, but a great many other choices will do, as well.) Then there exist at most countably many deterministic laws (if any) compatible with the facts in V – each determining a unique history for the rest of the universe – but uncountably many Humean possibilities to complete the mosaic.

Hence, we conclude that whatever the facts in V, it is atypical for the rest of the Humean mosaic to be constituted in a way that is consistent with a deterministic law (formulated in any language, formal or natural, that we could ever hope to understand).

As a Corollary, we obtain: Assuming Humean supervenience, any deterministic system that can describe a world up to time t will typically fail to be true at later times. This supports and strengthens the argument that Humeanism cannot sustain inductive inferences (Dretske, 1977; Armstrong, 1983). Of course, induction is difficult to justify in general, but Humean metaphysics *undermine* it in this sense.

While this (a)typicality result seems serious enough, it is, strictly speaking, a conditional claim "given one part of the mosaic." In general, there are already uncountably many possibilities for the "initial" data, i.e., uncountably many worlds consistent with every single deterministic law. At this point, we thus need some measure theory, after all, to obtain an unconditional typicality result. As always, we assume that one-element subsets (and hence, by σ-additivity, countable subsets) of an uncountable set have measure zero. In addition, we require only that this remains true if we conditionalize on the configuration of the Humean mosaic in V and count the possible configurations in some distant region U. This is certainly legitimate considering the Humean principle of free combinations which holds that one puts no restrictions on the other (our assumption is even weaker than that possible configurations in V and U can be measured independently). There is one technical subtlety involved in the proof which is given in the appendix (because we are potentially conditioning on a null-set), but in a nutshell, the argument concludes as follows: Denote by w_U the configuration of the mosaic in a spacetime region U. There are uncountably many possible configurations w_U, but (by the previous argument) at most countably many consistent with a deterministic law and the "boundary condition" w_V. Hence, $\mu(w_U \textit{ consistent with a law } | w_V) \equiv 0$ (for a suitable choice of U and V) and thus, with W^* the set of lawful Humean worlds,

$$\mu(W^*) \leq \int \mu(w_U \textit{ consistent with a law } | w_V) \, \mathrm{d}\mu(w_V) = 0$$

according to any reasonable measure. We conclude:

Theorem. *It is atypical for Humean worlds to be consistent with any deterministic systematization.*

Philosophically, the notion of a "reasonable measure" is doing a lot of work here. Mathematically, it is certainly possible to define other measures, but these are so clearly biased or ad hoc that they cannot play the role of a typicality measure. As we have said before, it is also possible, mathematically, to put a delta-measure on the reals and say that "almost all real numbers are zero." But this statement is only true in the technical sense in which the locution "almost all" is introduced in measure theory. In any other sense of the words, it is merely an abuse of language. The point is that a typicality statement will have rational (normative) implications if and only if it is made with respect to a reasonable notion of "almost all" (or "large" versus "small" sets of possible worlds). And the claim is that the assumptions of our theorem are so weak and well-motivated that they exhaust all measures that could pass for "reasonable." To deny our conclusion is to either deny that a one-element subset is vanishingly small compared to an uncountably infinite set (which seems absurd) or to *presuppose* extremely strong "correlations" between different parts of the mosaic (which means, in effect, to deny Humeanism).

Finite systematizations

Our proof that the existence of a deterministic systematization is atypical for Humean worlds relied on the assumption that there are uncountably many possible configurations of the mosaic, or an infinite number of physical facts that the laws would have to account for. On what basis could this assumption be denied? One could insist that the world is finitary, i.e., that space and time are finite and discrete, and that there are no continuous degrees of freedom in the physical ontology. While this cannot be ruled out in principle, it constitutes a very strong a priori commitment and a revisionary stance with respect to contemporary physics. Alternatively, one could insist that laws of nature do not have to provide a complete (microscopic) description of the world but only systematize a particular, limited subset of (macro-)events – e.g., measurement results or empirical observations – that is plausibly finite. This second option is essentially instrumentalism; the view that laws are efficient book-keepers of empirical data rather than universal truths about the world.

In any case, if laws had to account only for a finite number of physical facts, it would still be true that typical Humean worlds are more or less irreducibly complex – meaning that they cannot be systematized by laws that are significantly simpler than a complete list of the relevant events – but only with respect to a more restricted set of languages in which the systems can be formulated. (Think about the Chaitin model and the question, whether the Kolmogorov complexity of a finite sequence is significantly lower than the length of that sequence.) One could thus retreat to the idea that the order in our universe is not objective, but that (instrumentalist) laws – and the regular patterns they summarize – exist because we have adapted our cognitive and mathematical tools to the world that we inhabit (see, e.g., Wenmackers (2016)).

Although I find it very uncompelling, I am not going to argue against this possible escape. If one concedes that Humeanism is de facto instrumentalism (or requires revisionary physics) the whole debate would be a very different one.

Indeterministic laws

The issue becomes more complicated if we consider the possibility of indeterministic laws. Logically, at least, an indeterministic law (e.g., a stochastic evolution) could be compatible with any mosaic whatsoever – that is, unless there are real propensities in the world that the law is supposed to summarize. In fact, there is even a good case to be made that typical Humean worlds are well described by something like Brownian motion, which can be technically considered a "law" but describes pure noise rather than any kind of regular order.

For a probabilistic law to be *informative*, and allow for something akin to causal inferences, it must predict reasonably high conditional probabilities for a relevant class of events (Lewis (1980) talked, in particular, about "history to chance conditionals"), that is, expressions of the form $\mathbb{P}(A \mid B) \approx 1$ where the conditional probability for A depends non-trivially on B. In our world, the history of the universe up to the present time t should make it reasonably likely that the earth will still be in its solar orbit 10 seconds from now. Kicking a ball from the left/right should make it reasonably likely that the ball flies off to the right/left. More generally speaking, a concentration of masses in a small spacetime region B might make it very likely that masses agglomerate in another region A, or something like that.

Now, could such correlations be typical with respect to Humean ontological possibilities? I claim that they can not. If we take Humean metaphysics seriously, the possible configurations in one part of the mosaic should be independent of the facts in any other part of the mosaic. In effect, any evidence for a robust correlation is evidence that we do not live in a typical Humean world. And the existence of infinitely many correlated events would certainly be atypical with respect to Humean possibilities (while, if there is only a limited number of events that a law has to account for, we are essentially back in the "instrumentalist" scenario discussed above).

This is, admittedly, a less rigorous argument than the one for deterministic laws. And the result is weaker, as well, relying on a distinction between "informative" and "non-informative" laws that would warrant further elaboration, and on typicality measures with strong independence properties. However, at the end of the day, I don't expect the contentious point of our discussion to be whether Humean metaphysics fares much better with respect to probabilistic laws than deterministic ones. As with instrumentalism (or maybe even more so), committing to indeterminism from the get-go does not seem like an attractive option that most Humeans would want to take.

13.4 On the Uniformity of Nature

With the caveats just discussed, I consider it a fact that the existence of Humean laws is atypical among Humean worlds. The philosophically more delicate discussion happened in Section 3, where we argued for the normative implications of such typicality facts. I take it that any form of rationality is normative. At the same time, I see one of the weaknesses of probabilistic arguments in that they try to shortcut the issue and argue directly in epistemic or doxastic terms. Typicality facts are neither epistemic nor doxastic facts. They hold independently of what we know or believe. Yet, they have implications for what we *should* rationally believe, or accept, or seek to explain; in this case that we cannot accept Humean metaphysics and believe in a lawful universe without seeking explanation for its lawfulness.

On the other hand, the explanation provided by non-Humean laws is not a bona fide typicality explanation in that non-Humean laws make their instantiation not just typical but necessary. (Although, as discussed in detail in earlier chapters, they generally make macroscopic/empirical regularities typical but not necessary.) Technically, necessity implies typicality, but it is, of course, strictly stronger. The primary role of typicality in the argument is thus not to sustain an explanation but to establish that one is required, that the price for declaring the history of the universe to be entirely contingent is unreasonably high. However, what applies here as well as to bona fide typicality explanations is that they do not have to involve an interesting "mechanism" by which the explanandum comes about. There is no interesting story left to tell about how laws govern or how dispositions bring about their manifestation; the point is that they are a natural part of an ontology that doesn't make the existence of regularities in the world miraculous.

This explanatory virtue of non-Humean laws comes from their modal force, from the way in which they restrict ontological possibilities. In contrast, the idea that non-Humean laws fare better in explaining their particular instances has made the modal realist positions vulnerable to the *virtus dormitiva* objection that any explanation they provide over and above the regularity theory is trivial or circular: Why do masses attract each other? Because they have the disposition to attract each other. Or: because it is a law that masses attract each other. In the contemporary literature (see, e.g., Emery (2019)), such statements are often spelled out in terms of grounding relations or as "in virtue of" explanations, which makes them manifestly non-circular but still ring hollow to people not already sold on the merits of these metaphysical concepts. Indeed, the impactful argument of Loewer (2012b) – which not only rejects the charge that Humean laws are not explanatory but puts anti-Humeans on the defensive – was to insist on a distinction between scientific and metaphysical explanations, suggesting that the latter are ipso facto unscientific and thus somehow suspect. Thinking in terms of typicality (which, we have argued, is very scientific) one understands that the actual explanatory advantage of non-Humean laws is not that they provide an additional metaphysical ground for individual instances, but that they account for why our world is lawful in the first place.

At the end of the day, one can only go so far in compelling someone to accept a particular way of reasoning and the norms that come with it. Some readers may deny that typicality facts have any philosophical implications at all, that there is even a sense in which Humean metaphysics make the lawfulness of our world surprising or remarkable. There is, however, no shame in sharing at least in a sense of wonder about the order of our cosmos. (After all, according to Aristotle, the sense of wonder is the very beginning of philosophy.) The following passage from one of Albert Einstein's letters to Maurice Solovine comes to mind:

> You find it strange that I consider the comprehensibility of the world (to the extent that we are authorized to speak of such a comprehensibility) as a miracle or as an eternal mystery. Well, a priori one should expect a chaotic world which cannot be grasped by the mind in any way. One could (yes *one should*) expect the world to be subjected to law only to the extent that we order it through our intelligence. Ordering of this kind would be like the alphabetical ordering of the words of a language. By contrast, the kind of order created by Newton's theory of gravitation, for instance, is wholly different. Even if the axioms of the theory are proposed by man, the success of such a project presupposes a high degree of ordering of the objective world, and this could not be expected a priori. That is the "miracle" which is being constantly reinforced as our knowledge expands. There lies the weakness of positivists and professional atheists who are elated because they feel that they have not only successfully rid the world of gods but "bared the miracles." (Cited from Einstein (1987, pp. 132-33).)

What, to their credit, distinguishes most Humeans from the "positivists and professional atheists" that Einstein talks about, is some acknowledgment that the best system account of laws has to rely on nature being "kind to us" (Lewis, 1994, p. 479), on "a high degree of ordering of the objective world" that cannot, by any means, be expected a priori. However, this kindness of nature is *so* stupendous and is doing *so* much work in the best system account that it is highly unsatisfying, if not intellectually dishonest, to leave it as an afterthought or some sort of auxiliary assumption without any basis in the metaphysical theory. If Humeans tried to give it more flesh, and spell it out as a metaphysical principle that makes the uniformity of the world typical (or necessary)[4], their account would be much more sound but also start to look a lot more like anti-Humeanism.

One the other hand, some authors have made the point that anti-Humean metaphysics fare no better in explaining the uniformity of nature (Hildebrand, 2013). In this vein, advocates of the regularity theory could admit that Humeanism fails to account for a lawful universe but deny that anti-Humean positions have an explanatory advantage in this respect. In the language of typicality, the relevant argument goes roughly as follows:

Even if our world contained primitive laws or dispositions that necessitate simple universal regularities, this very fact is atypical, as well. In almost all worlds in which

[4] A metaphysical analog of the Past Hypothesis in physics, so to speak.

non-Humean laws exist, these laws would be so strange or complex that they could not be formalized in any simple and informative system. Hence (it seems), the typicality argument can be turned just as well against the anti-Humean theories.

I am not sure if this (a)typicality statement is true. At least, most anti-Humean theories do not entail the possibility of arbitrarily complex laws in the same sense in which Humean metaphysics entails the possibility of arbitrarily complex mosaics. Moreover, if we change the configuration of a lawful Humean mosaic only slightly (in a small spacetime region, let's say,) it will, in general, no longer be a lawful mosaic. If we change a simple law only slightly, it will still be a simple law. The point is that the "degrees of freedom" of a law are clearly different from those of the world, and the question, what metaphysical possibilities we must admit with respect to the type "law of nature" strikes me as a very difficult one. Hildebrand (2013) takes nomic primitivism to mean that there exists a primitive lawhood operator "It is a law that...", which can attach to any proposition P, no matter how gruesome or unnatural. But this is not at all how physical laws are formulated, or what the anti-Humean theories that we regarded as promising actually commit to.

That said, even if we grant that typical non-Humean worlds have no simple laws, it is crucial to note that this is a typicality statement with respect to a different reference class than we employed in our discussion; namely metaphysically possible worlds – under a liberal interpretation of metaphysical possibility – rather than what we called ontologically possible worlds. It is thereby shifting the debate from ontology to meta-ontology, from the question: "What is the fundamental ontology of our world (and does it contain the laws that physics discovers)?" to: "Why is the fundamental ontology (here, specifically, the laws) what it is?". It is much less clear that this is a good and tractable question, and it is, in any case, not the question we set out to debate. It might be worth exploring the idea of meta-laws that constrain the possible non-Humean laws (Lange, 2009), but this goes beyond the scope of our discussion, and one must worry that it would, at best, be passing the buck (for what explains or necessitates the meta-laws?).

The following analogy may help to illustrate my point: If all matter propagates along three spatial dimensions (not just appears to, but actually does), it is more than reasonable to infer that space, however conceived, *is* three-dimensional. (It is possible yet atypical that space has more dimensions, while all motion happens to occur along a three-dimensional subspace). But why has space three dimensions when it could, at least mathematically, have arbitrarily many? I don't know, and this was not the issue.

The aim of our discussion was not to defend anti-Humeanism as an a priori thesis. No one, I think, holds the view that our world must contain some primitive laws or dispositions, even if they govern only the growth of beetroots, or account for no meaningful regularities at all. My belief in non-Humean laws is very much contingent on the success of the scientific enterprise. And if I wake up tomorrow and find that the law of gravitation no longer holds, I would float through the air and admit that Humeanism was probably right all along.

Certainly, anti-Humean metaphysics do not relieve us of wonder and amazement about the simple and elegant laws that we discover in our universe. The existence of something over and above the Humean mosaic is, instead, an ontological conclusion that we draw from this discovery – with good reason, as this chapter has argued in detail. That may be as far as we can go. However, if there were a promising chance to take the explanation one step further, to understand *why* the laws are what they are, we should, by all means, follow the evidence where it leads us. It could, in any case, lead us only further away from Humeanism.

Appendix: Proof of the Theorem

Theorem. *It is atypical for Humean worlds to be consistent with any deterministic systematization.*

Proof. Let's assume, with David Lewis, that the fundamental ontology is one of "perfectly natural properties" instantiated at spacetime points. (The argument for other ontologies, e.g., continuous particle trajectories, will go more or less analogously.) We can then model the set of possible Humean worlds by $W := \{w : \mathcal{M} \to S \subset \mathbb{R}^n\}$, where $\mathcal{M}$ is the spacetime manifold and the "field values" $w(x)$ describe the magnitudes of the relevant properties at spacetime point x.

We denote by w_V the restriction of w to $V \subset \mathcal{M}$ for a suitable V as explained in Section 4.2 (w_V is "the configuration of the mosaic in V") and by $L_U[f]$ the possible configurations of the mosaic in $U \subset \mathcal{M}$ that are consistent with some deterministic law and the boundary condition $w_V = f$. By the argument given in Section 4.2, $L_U[f]$ is at most countable for any $f : V \to S$, and for any $U \subseteq \mathcal{M} \setminus V$, the set W^* of Humean worlds consistent with a deterministic law must be contained in $\{w \in L_U[w_V]\} \subset W$.

Now, we choose as U a collection of points in $\mathcal{M} \setminus V$; countably infinitely many points if S is discrete, and finitely many if S is continuous. In any case, there are uncountably many possible configurations on U (but at most countably many consistent with a deterministic law and given boundary conditions on V). Let μ be a normalized measure on W (more precisely, on a suitable σ-algebra). Then there exists a regular version of $\mu\left(w(U) \in \cdot \,|w_V\right)$, i.e., a well-defined measure on the possible configurations in U, even if we conditionalized on a null-set. This holds because the value space of w_U, viz. $S^{|U|}$, is isomorphic to some subspace of $\mathbb{R}^k$, $k \in \mathbb{N} \cup \{\infty\}$ (Ash and Doleans-Dade, 2000, Thm. 5.6.5). By assumption, this conditional measure has no discrete part (for at least some suitable choices of U and V), i.e., it is zero on singletons, and thus by σ-additivity also on countable sets. Hence, $\mu(w_U \in L_U[w_V] \mid w_V) \equiv 0$. Therefore,

$$\mu(W^*) \leq \int \mu(w_U \in L_U[w_V] \mid w_V)\, \mathrm{d}\mu(w_V) = 0. \tag{13.1}$$

The proof extends at least to σ-finite measures with the conditional "probability" replaced by a Radon-Nikodym density (Ash and Doleans-Dade, 2000, Thm. 2.2.1).

□

Chapter 14

Super-Humeanism: A Starving Ontology

14.1 The Radicalization of Humeans

Ever since its inception, which is widely attributed to the work of David Lewis (see, e.g., Lewis (1973, 1986a, 1994)), the program of Humean supervenience has grown bolder and larger in scope. While Lewis's account of laws of nature was mainly motivated by his rejection of modal connections, other authors have employed the best system strategy to ban all sorts of properties or entities, which they deem metaphysically suspicious, from the fundamental ontology of the world. Indeed, from the point of view of many modern Humeans, Lewis was too generous in outfitting the Humean mosaic, by admitting intrinsic qualities in addition to spatiotemporal relations. The key to understanding the more recent and ambitious Humean programs is to appreciate the following two insights:

i) All empirical data can be ultimately understood as consisting in the distribution of matter in space and time (including pointer positions, display readings, computer printouts, or whatever else records the outcomes of experiments).

ii) While the strength of the best system candidates is measured against the regularities in the world, the laws of nature, as described by our best physical theories, are not just universal generalizations but involve all sorts of physical constants, dynamical variables, and geometric structures that figure in the mathematical formulation of laws of temporal evolution.

Hence, the Humean, if she is bold enough, can maintain that the Humean mosaic consists only in the spatiotemporal distribution of matter — made up by localized objects such as particles — while all other structures and quantities appearing in the formulation of physical theories are part of the best system, introduced to provide a simple and informative summary of the mosaic.

Structures and quantities that have thus been subject to some form of Humean reductionism include physical properties such as mass and charge (Hall, 2009; Esfeld et al., 2015), dynamical objects like the wave function in (Bohmian) quantum mechanics (Esfeld et al., 2014; Esfeld, 2014b; Miller, 2014; Callender, 2015; Bhogal and Perry, 2017), and even geometric structures describing an absolute space or spacetime (Huggett, 2006; Vassallo and Esfeld, 2016). This is to say that the Humean can defend a relationalist conception of space (and time), without providing a relational reformulation of the physical laws, by maintaining that the absolute background space, presupposed by theories such as Newtonian mechanics or General Relativity, is merely a descriptive tool, allowing for a more efficient summary of the relational history of matter. The corresponding view is also known as *Super-Humeanism*, in analogy (and opposition) to super-substantivalism in the philosophy of spacetime.

Of course, not all Humeans go equally far in their reductionism, and not all reductionist programs are based on Humean metaphysics. Being anti-Humean doesn't mean that one has to be a realist (or "fundamentalist") about every structure employed in our mathematical formalization of the laws. However, the best system strategy, once applied to laws of nature, has proven very efficient in ridding the world of "non-Humean whatnots" and thereby opened the door to declaring anything one likes – or rather dislikes – a *whatnot*, deserving of the Humean exorcism.

One of the most radical and comprehensive Super-Humean projects is due to Esfeld, Deckert and Vassallo (e.g., Vassallo et al. (2017); Esfeld and Deckert (2017)). The aim of this project is to spell out an ontology of the natural world that is as parsimonious as possible, while being overall coherent and empirically adequate. The proposed ontology consists in primitive matter points, without intrinsic properties, individuated by primitive distance relations (axiom 1). These matter points are permanent, with the distances between them changing (axiom 2).

All that exists in the world is thus a network of changing distance relations between bare matter points. The role of physics is now to provide the most efficient description of the history of changing distance relations – striking an optimal balance between simplicity and strength – by introducing appropriate dynamical parameters and geometric representations. In a recent book (Esfeld and Deckert, 2017), this view is spelled out in detail and applied to a wide range of modern physical theories from Newtonian mechanics, to Bohmian quantum mechanics, to general relativity and even quantum field theory, where the authors show how an ontology of permanent point particles can account for the phenomena appearing as "particle creation and annihilation" (cf. Deckert et al., 2017).

It is important to note that the distance relations postulated by Esfeld and collaborators are structureless, dimensionless, and undirected. They are constrained only by the triangle inequality which is arguably the minimal requirement to make them distance relations that hold together the material world (in contrast to, let's say, hypothetical thinking relations that hold together a world of minds), but there is nothing about the relations that would make them 3 or 2 or 10-dimensional, Euclidean or non-Euclidean distances. This is in notable contrast to the relational project championed, in particular, by Julian Barbour (see, e.g., Barbour and Bertotti (1982)), whose fundamental relations carry much more geometric structure (precisely the non-absolute Euclidean structure in the non-relativistic case) and are meant to constrain the dynamical laws, in the sense that the formulation of the laws should not refer to more or different geometric structure than what is provided by these fundamental relations. Upon the Super-Humean strategy, first proposed by Nick Huggett (2006) and adopted by Esfeld et al., all geometric features have been delegated to the best system. The Humean mosaic consists only in structureless distance relations between point particles (primitive matter points), while the candidates for the best system description vary not only with respect to the dynamical laws but also with respect to the spatiotemporal geometry used to represent the history of distance relations and formulate the laws of motion. Hence, from the fundamental point of view, all geometric structures are nothing more than mathematical or representational surplus. Yet, of all the representational surplus that one could posit, one combination (the Super-Humean hopes) will be “true” in virtue of striking the optimal balance between simplicity and strength. In this sense, the (apparent) geometry of space or spacetime supervenes, together with the dynamical laws, on a purely relational, non-modal, and non-geometric ontology.

We should pause for a second to appreciate the implications of this view. It is to say that space appears to be 3-dimensional, not because geometric relations instantiated in the world are 3-dimensional, but because the contingent history of distance relations happens to be such that it can be most efficiently summarized in a 3-dimensional representation. It is also to say, for instance, that when Einstein presented his theory of relativity, he didn’t discover anything new about the nature of space and time, but rather observed that the history of distance relations could be more efficiently summarized in a 4-dimensional Lorentzian spacetime geometry. Ultimately, according to the view of Esfeld, Vassallo, and Deckert, physics seems neither capable nor in charge of informing us about the fundamental ontology of the world. Since they have established that a) all empirical facts can be understood as facts about the configuration of matter and b) facts about the configuration of matter can be conceived as facts about distance relations between primitive matter points, it is unclear what empirical evidence or scientific discoveries would compel them to revise their ontology.

Thus, I believe that by taking parsimony and Humean reductionism to an extreme, Esfeld et al. actually demonstrate that Super-Humeanism as an a priori thesis is both uninteresting and uncompelling since it is too promiscuous and universally applicable. When all is said and done, there is one and only one line of reasoning that is used to ground successful physical theories in her preferred ontology. I call it the *Super-Humean subterfuge*:

> Our physical description of the world exhibits the feature X because the contingent, relational distribution of matter throughout the history of the

> universe happens to be such that the best system description exhibits the feature X.

As suggested before, the problem with such reasoning is not that it couldn't be coherently defended, but that the arguments put forward in defense of Super-Humeanism would seem to apply independent of what X stands for. If every part of a physical theory can be delegated to the best system mythology, no part needs to affect their ontological commitments.

Notably, the claim here is not that a priori everything could supervene on regularities in the relational distribution of matter. Qualia or normative facts almost certainly don't – but then they are also not within the scope of any physical theory. The basic idea underlying Esfeld's project is rather that a Humean mosaic made up of relational configurations of matter is sufficient to ground all empirical facts that are the target of naturalistic accounts; and that any other structure appearing in the formulation of a physical theory can then be understood through its role in describing or summarizing such a mosaic (or else it is superfluous).

In this sense, the metaphysics of Esfeld et al. is committed to naturalism (or even physicalism) but it is not, despite their claims, naturalized metaphysics in the sense of being guided or informed, in any significant way, by our best scientific theories. This may seem like a mere methodological critique, unless we already took for granted that only naturalized metaphysics is good metaphysics. However, a reasonable standard for rejecting their proposal cannot be to show that it is strictly impossible or inconceivable as an ontology of the natural world (I grant that it is not). Instead, focusing on Super-Humeanism in the narrower sense as an interpretation of space or spacetime, I am going to argue that structureless distance relations as a ground (or supervenience basis) of a 3-dimensional geometry of space, respectively a 4-dimensional geometry of spacetime, must be rejected on the basis that it makes the latter atypical.

14.2 Space, lost

The Super-Humean strategy to ground physical theories in a minimal relational ontology was laid out by Nick Huggett (2006) in his relationalist (re)interpretation of Newtonian mechanics. This will also serve as the test-case for our following discussion. While Newtonian mechanics is no longer considered a fundamental theory, it is still the preferred playground of many philosophers; and while it may not provide the best system description of the world, it is still a very good and useful one. A possible extension of Huggett's regularity account to general relativistic spacetime is discussed in Esfeld and Deckert (2017, Ch. 5) as well as in Vassallo and Esfeld (2016). By and large, it is just an adamant application of the Super-Humean subterfuge; the following objections will carry over accordingly.

I want to emphasize that my objections are targeted against the full-fledged Super-Humeanism spelled out by Huggett and adopted by Esfeld et al. The aim of their regularity theory is essentially two-fold (the second being an escalation of the first):

1. To ground physical theories referring to an absolute background space or spacetime in a relationalist ontology without providing a relational reformulation of the physical laws.

2. To reduce all geometric features of this background space – its dimension, symmetries, curvature etc. – to thin distance relations that carry no instrinsic geometric structure.

These two points are not entirely independent: the more structure the fundamental relations carry, the less flexible will the Super-Humean ontology be in accommodating different background spaces described by different theories. In principle, a not-quite-so-super Humean could let go of the second aim while still invoking Huggett's regularity theory to accomplish the first, i.e., postulate thicker spatial or spatio-temporal relations as the supervenience base of the absolutist physical description.[1] Our following discussion is primarily concerned with point 2, that is, with the Super-Humean claim to get away without committing to bona fide geometric structures in her fundamental ontology. In other words, the conflict here is not between relationalism and substantivalism, but between a geometric and a non-geometric ontology, making the substantivalist, the spacetime structuralist, and even the liberal relationalist à la Barbour or Saunders (2013) unlikely allies against the Super-Humean minimalism.

To make precise what the Super-Humean relations are, we recall the definition of Esfeld and Deckert (2017), who specify the relational structure instantiated by N matter points by the following set of axioms. (For convenience, the matter points are labeled by an index $i \in \{1, \dots, N\}$; the indices are, however, arbitrary and not supposed to indicate a primitive identity of the matter points.)

i) Any two matter points stand in a distance relation that can be represented by a positive real number $r_{ij} \in \mathbb{R}^+, 1 \leq i \neq j \leq N$.

ii) The relation is symmetric, i.e., $r_{ij} = r_{ji}, \forall 1 \leq i < j \leq N$.

iii) The numerical assignments satisfy the triangle inequality: $r_{ik} \leq r_{ij} + r_{jk}$.

Esfeld and Deckert also require that the matter points are individuated by the distance relations (thereby insisting on what has become known as absolute discernibility). That is, the matter points have no intrinsic identity, but any two points are numerically distinct by virtue of standing in a different relation to at least one other matter point. To this end, the authors must add a fourth postulate, excluding symmetric configurations in which some particles would no longer be absolutely discernible. While this combination of relationalism and moderate ontic structural realism comes with its own set

[1] I don't want to go into a deep metaphysical discussion about thin versus thick relations; I am merely using the terminology to distinguish between the minimalistic relations of Huggett and Esfeld, that are constrained only by the triangle inequality, and other relational accounts in which the fundamental relations carry more geometric structure, e.g., a 3-dimensional direction in addition to a distance.

of interesting problems (most notably: what provides for the identity of the matter points *over time*, as the configuration of distance relations changes?), those are beyond the scope of this discussion. What matters for our purposes is that the above axioms are very weak, essentially emulating the general mathematical definition of a "metric." To repeat: fundamentally, there is nothing about the distance relations that would make them 3-dimensional or Euclidean, or put any other constraints on the dimension, curvature or topology of the physical geometry.

According to the Super-Humean regularity theory, the fact that Newtonian mechanics is formulated on a 3-dimensional Euclidean space only means that this geometric representation – together with the laws of Newtonian mechanics – strikes a good balance between simplicity and strength in summarizing the history of structureless distance relations. The candidates for the best system thereby vary not only with respect to the dynamical laws, but also with respect to the background space (qua mathematical structure) used to represent or embed the history of distance relations in the first place. Such an embedding has to preserve only the distances between the matter points since there is no other structure to preserve. That is, if we denote the history of distance relations by $r_{ij}(\lambda)$, with an arbitrary "time" parameter λ, the representation of the matter points as trajectories $q_i(\lambda)$ in some metric space (M, d) must satisfy $d(q_i(\lambda), q_j(\lambda)) = r_{ij}(\lambda)$. All other things being equal, a lower-dimensional space M and higher degree of symmetry are preferred on the grounds of simplicity (Huggett, 2006, p. 54), but the fact that space appears to be 3-dimensional, and at least locally Euclidean, is purely contingent according to the Super-Humean account.

I agree with Belot (2000, p. 10) that such an interpretation "is arrant knavery: a cheap instrumentalist rip-off of Newtonian theory," but I doubt that one can shame the Super-Humean into yielding any ground. Instead, I will make the case that the Super-Humean reduction of Euclidean space is, by reasonable standards, empirically inadequate.

What the account fails to acknowledge, is that it is no trivial matter to embed such a relational network into 3-dimensional Euclidean space. It is always possible to embed a triplet of particles: the result is simply a triangle (or a line, in the degenerate case) whose side-lengths correspond to the distances between the matter points. The triangle inequality – which the distance relation must satisfy by definition – is sufficient to ensure the existence of such an embedding. However, in order to embed four, or five, or six matter points, additional constraints on the distance relations would have to be satisfied, which would be purely accidental upon the Super-Humean view. A similar observation was made by (Maudlin, 2007a, pp. 87-89), who argues that the substantivalist, but not the relationalist, can explain such constraints, in particular the basic triangle inequality that the relationalist has to accept as an axiom (while the substantivalist can derive it from the concept of "path lengths"). I am not concerned about the status of the triangle inequality but want to elaborate on the case of multiple particles, that is, on the additional constraints that are not even axioms but mere contingencies upon the Super-Humean account. On this basis, I will defend the thesis that the most basic empirical fact about space, namely its low dimensionality, is not just unexplained by the regularity theory but should be regarded as a falsifying instance.

To illustrate why the embedding of multiple matter points into 3-dimensional Euclidean space requires additional constraints, assume we got lucky and were able to embed four matter points while preserving their mutual distances. Embedding a fifth matter point now corresponds, pictorially speaking, to determining the point of intersection of four spheres in 3d-space (each centered around one of the existing matter points with radius equal to its distance to the fifth). However, four spheres in 3-dimensional Euclidean space do not have a point of intersection, unless their radii happen to satisfy additional algebraic relations. In general, we would have to move into 4-dimensional space to faithfully represent the distance relations between all five matter points. And the larger the network of distance relations, the more (and more specific) constraints on the relations must be satisfied to fit a given configuration into a 3-dimensional geometry.

More precisely, embedding N points into d-dimensional Euclidean space amounts to assigning $d(N-1)$ coordinates. (Without loss of generality, we can fix one particle to the origin of the coordinate system, leaving $N-1$ points with d coordinates each. In fact, accounting for global rotations, we have only $d(N-1)-\frac{1}{2}d(d-1)$ relational degrees of freedom – and one less if we factor out absolute scale – but these corrections are negligible for the following estimates.) On the other hand, there are $\binom{N}{2} = \frac{1}{2}N(N-1)$ distance relations between N matter points, amounting to the same number of quadratic equations for the coordinates as we want to realize the relations as Euclidean distances on $\mathbb{R}^d$. (The coordinates of two points $x, y \in \mathbb{R}^d$ with distance r must satisfy the quadratic equation $(x_1 - y_1)^2 + \ldots + (x_d - y_d)^2 = r^2$.) In conclusion, as soon as $N > 2d$, the system is overdetermined and will not admit any solution at all unless roughly $\frac{1}{2}N(N-1) - d(N-1) = (\frac{1}{2}N - d)(N-1) \approx \frac{1}{2}N^2$ additional constraints are satisfied.

For our universe, N is estimated to be about 10^{80}. Hence, the number of geometric constraints that would have to be satisfied – purely accidentally – for a network of Super-Humean relations to be embeddable into 3-dimensional Euclidean space is at least of the order 10^{160}. Yet, the Super-Humeans ask us to believe that this epic coincidence obtains not only within one particular configuration, but that the trillions of trillions of trillions of constraints happen to be preserved over time, as the distance relations between the matter points change. Mind you, according to their metaphysics, there is nothing in the ontology that would make it necessary for these constraints to be satisfied in any configuration, let alone propagate with the dynamics.

To put it the other way around: by assuming a 3-dimensional space in which the matter points move – that is, a real physical space that constraints their motion, not just a mathematical construct that describes it – we explain about 10^{160} dynamical constraints that the Super-Humean accepts as bare facts. This is a successful reductive explanation if there ever was one.

Let's phrase these observations as a proper typicality statement. The ontologically possible Super-Humean configurations form an abstract space of dimension $\approx \frac{1}{2}N^2$. The configurations that could be embedded into a d-dimensional Euclidean space form a submanifold of dimension $\approx dN$. For $d \ll N$, such configurations are clearly atypical, most obviously in the dimensional sense and then also in the sense that they form a set of measure zero with respect to any absolutely continuous measure. The same holds (modulo some measure-theoretic subtleties) if we consider ontologically possible worlds, i.e., trajectories in the respective configuration spaces rather than instantaneous configurations. For all the reasons discussed in the previous section, Super-Humeanism must, therefore, be rejected as a metaphysical account of the low-dimensional space or spacetime that we actually live in.

I imagine a Super-Humean response to go somewhat like this: Admittedly, we cannot give a deeper metaphysical explanation for why the history of distance relations is such that it can be represented on a 3-dimensional Euclidean space. But you, as a "substantivalist," cannot give any deeper metaphysical explanation for why space is 3-dimensional (and at least locally Euclidean), rather than 4 or 5 or 17,000 dimensional. Of course, the 3-dimensional representation of the world contains a lot of information about the history of distance relations, but this is precisely what makes this geometry part of the best system. In other words, all those algebraic relations that hold between the particle distances are part of the regularities that we find in the world, and they are summarized in our 3-dimensional representation of the (Newtonian) laws, which thus provides the kind of unifying explanation that we should expect from science.

Prima facie, this Humean response may have some persuasive power, but it is based on a completely false equivalence. Obviously, a substantivalist or a liberal relationalist (who admits thicker relations as fundamental) postulates an absolute space, respectively some geometric structure, as a physical and metaphysical primitive. She thereby accepts certain facts about the geometry of the world as primitive, such as the fact that space has 3 dimensions rather than 4 or 5 or 17,000. But every theory needs some primitives. And the geometry of space or spacetime is arguably so basic to our conception of the world, and so informative about the possible configurations of matter in it, that it is a more than reasonable choice as an ontological primitive. The Super-Humean doesn't deny that space or spacetime are real in some sense – that there are true geometric facts about the world – but she claims to provide a reductive account of these facts. It is she who must, therefore, deliver on her promises. The Super-Humean reduction, however, fails since it has to assume an exceedingly special configuration of distance relations throughout the history of the universe to account for even the most basic geometric facts.

One reason why we should not accept such an account as successful is that it "reduces" a small number of simple primitive facts to a huge number of complicated primitive facts. One reason why we cannot accept such an account as successful is that we would give up on any means to test a theory about what exists in the world against any facts that we could possibly know about the world if we admitted that the actual world need not look anything like a typical one that the ontology can possibly constitute.

14.3 Empirical Adequacy in Metaphysics

Against this backdrop, I claim that the Super-Humean account of space is not just explanatorily deficient but empirically inadequate. Remember that when embedding N points in a d-dimensional Euclidean space, we have the freedom to choose $d(N-1)$ coordinates, while having to solve for $\frac{1}{2}N(N-1)$ quadratic equations expressing the distance relations between the matter points. This means that given a typical set of distance relations, the dimension of the lowest-dimensional space into which it can be embedded will be of the order $d \approx \frac{1}{2}N$, with $N \approx 10^{80}$! In this sense, the metaphysical theory of Super-Humeanism makes a prediction, namely that space should look extremely high-dimensional (if it has any meaningful geometric structure at all). Since the opposite is true, and we seem to live in 3-dimensional space or 4-dimensional spacetime (even the 10- or 11-dimensional spaces required by string theory would count as very low-dimensional in this context), we must conclude that a Super-Humean ontology of structureless distance relations doesn't fit the world that we experience.

I have, of course, raised an analogous objection against the regularity view of laws, showing that a typical Humean world has no law-like regularities in the first place. But the non-Super-Humean can at least with some credibility carry the banner of empiricism. She can argue that she and her opponent agree on all concrete physical facts in the world and disagree only about the status of modality. And this allows her to make the case – though ultimately unsuccessful – that since we have no direct empirical access to modal relations, we should go with the deflationary, i.e., more parsimonious account. The Super-Humean, however, is in an even worse position with respect to space or spacetime. She and her opponent – be it the substantivalist or the liberal relationalist à la Barbour – do *not* agree on the concrete physical facts in the world. For her opponents, the spatial relations instantiated in the world (be it fundamentally between particles or between particles in virtue of occupying certain points in space) carry more geometric structure, characterized not only in terms of distances but also in terms of 3-dimensional *directions*. And it is these thick relations – not the thin relations – to which we have the most direct empirical access. We experience objects as having a shape and a location – not only a distance – relative to one another. Hence, even the Super-Humean must admit that it is the 3-dimensional representation, not the unordered list of primitive distance relations, that has the most intimate connection to our manifest image of the world. And then the Super-Humeans can call their relations “spatial” all they want. What they have in their ontology is not space, as commonly understood and experienced, but a much more impoverished notion; starving thin relations that can only emulate spatial relations for the price of assuming an utterly atypical configuration throughout the history of the universe.

To sum up more systematically, let us return to the promiscuousness of the Super-Humean subterfuge discussed before. That a proposal for a fundamental ontology of the natural world can, in principle, fit our experience and scientific description is an extremely low bar that saves us neither from arbitrariness nor from absurdity. In principle, one can postulate any ontology, provided that it affords enough degrees of freedom to tweak, and assume that these degrees of freedom are arranged in precisely such a way as to instantiate (or be in some sense isomorphic to) whatever structures are identified in nature. The question we must ask is therefore: what distinguishes a serious and plausible candidate for ontology from a mere metaphysical prejudice? Or, in other words: what distinguishes a legitimate application of the subterfuge from one that is trivial, spurious, and ad hoc? A few possible criteria come to mind, and Super-Humeanism fails all of them:

1. We can require that the fundamental ontology matches, in some sense, the structures or objects appearing in the formulation of our best physical theories. This criterion – marking the still dominant methodology in naturalized metaphysics in the Quinean tradition – is explicitly rejected by Super-Humeans in general and Esfeld et al. in particular. Hence, it is not surprising that their account fails in this respect: no successful physical theory is formulated in terms of structureless distance relations between bare matter points, and arguably none ever will be.

2. We can require that our actual world resembles not just a particularly special and fine-tuned model of our metaphysical theory but a typical one. Humeans, in general, must reject this criterion, and Super-Humeanism fails it in a particularly spectacular fashion since a typical world formed and held together by structureless distance relations would not look anything like a material world in a low-dimensional space or spacetime.

   Note, on the other hand, that for a substantivalist, the apparent truism that “space looks 3-dimensional because it actually is 3-dimensional” involves just such a typicality reasoning. It is possible – though highly atypical – in a 3-dimensional Newtonian universe that all particles move on a 2-dimensional hyperplane. Such a world would appear 2-dimensional to the hypothetical flatlanders living in it. It is equally conceivable and physically possible that we are 3-dimensional “flatlanders” living in a higher-dimensional Euclidean space. The postulate of a 3-dimensional space matches the appearance of a 3-dimensional space not because it makes the latter necessary but because it makes it typical.

3. We can require that the connection between ontology and experience is, in some sense, direct, simple, and robust. In other words, this is to insist on a reasonably

> short cognitive distance between our manifest and scientific image of the world. (The terms "manifest image" and "scientific image" go back to Sellars (1962); for an interesting discussion along these lines, see Maudlin (1997)). Some version of this criterion is, in fact, the main argument for primitive ontology theories, which postulate a local ontology in 3-dimensional space or 4-dimensional spacetime, against functionalist approaches such as wave-function realism that try to connect our experience of the world to a more abstract description of physical reality by some sort of functionalist emergence.
>
> Esfeld and collaborators attempt to carry the banner of the primitive ontology program. Their terminology can easily suggest that the "Leibnizian distance relations" are just the intuitive spatial relations that we perceive between macroscopic objects, and that the objects themselves can be straightforwardly conceived of as a collection of primitive matter points. For the reasons just discussed, this would be quite misleading. The connection between the network of distance relations and our experience of the material world is not direct. It is provided in terms of best system representation of the history of distance relations, not in terms of the fundamental ontology. The connection is not simple. It must take the detour of the best system account, a highly complex procedure of trying and comparing different summaries and representations of the entire history of the universe. And the connection is not robust. It relies on an extremely special and meticulously fine-tuned arrangement of the Humean mosaic, and even a small perturbation of this arrangement would lead to a radically different appearance of the world.

The criteria for ontology that Esfeld et al. announce to follow is "parsimony – together with empirical adequacy." However, as we have just seen, the Super-Humean account is empirically adequate only by standards that render the criterion itself trivial; that is, in the sense in which any ontology could be postulated as the supervenience base of our scientific description of the world if it is just sufficiently complex and fine-tuned, and if the supervenience relations are sufficiently indulgent (as Humean supervenience is). Which leaves us with parsimony as the only success of their "minimalist ontology."

This is not good enough. Humeans who are guided by a principle of parsimony seem to believe that postulates about *what* there is in the world are costly, while assumption about *how* it is come for free. But the two issues cannot be disentangled since we cannot pass rational judgment on one without the other. Parsimony may be a good criterion for the *what* question, but we need additional criteria, such as the ones proposed above, to judge it in conjunction with a *how* hypothesis. To postulate a parsimonious ontology and then assume that this ontology is arranged however it needs to be to get the phenomena right is too cheap to lend any credibility to the ontological claims. And delegating the how questions entirely to physics, which should inform us how the fundamental entities are arranged in space and time, doesn't work either, unless we take our ontological hints from science, as well. Physics per se doesn't inform us about Super-Humean relations because no established physical theory is about Super-Humean relations.

The broader philosophical point, of which I hope to have convinced the reader, is that we must assess the empirical adequacy of a metaphysical theory by sufficiently robust standards – or else the criterion itself becomes trivial. *Typicality* is one of the standards I have proposed. Our world must match, in the relevant respects, a typical (or at least not atypical) model of the ontology. As argued in earlier chapters, this is to follow the good example of natural sciences, if not a necessity of thought.

The other standard was what I described as a *reasonably short cognitive distance between the fundamental ontology and our manifest image of the world.* This is not quite the same as saying that the ontology itself is intuitive. What we must be able to intuitively grasp is how the theory ultimately connects to the world that we experience, even if it describes a fundamental reality that diverges radically from our experience. At *some* point, manifest and scientific image must meet closely enough that the remaining gap can be easily jumped by our intellect. This is where the wave function or quantum state functionalism of Everettian quantum mechanics failed, and Esfeld's "bare matter point functionalism" fails in a similar fashion.

Esfeld and collaborators deserve much credit for pushing the Humean project to its limits, following the principle of parsimony summarized in Frank Jackson's armchair metaphysics credo that the methodology "is not that of letting a thousand flowers bloom but rather that of making do with as meagre a diet as possible." (Jackson (1994, p. 25), requoted in Esfeld and Deckert, 2017). Unfortunately, the proposed "minimalist ontology" is not just meagre but starving. It is isolated from any input from empirical sciences and fails as a metaphysical foundation of space or spacetime, both as given to us by basic intuition and as described by our most successful physical theories. In the end, the radicality of Esfeld's program demonstrates that while it may be true that we don't have direct empirical access to intrinsic properties, or necessary connections, or the geometry of spacetime, renouncing all of them at once leaves us with an ontology that is too impoverished to match the world that we experience.

Chapter 15

Special Science Laws

In this chapter, we will discuss the special sciences, in particular the reduction of special science laws to fundamental physical laws as typical regularities. We will focus mostly on the example of biology because more specialized sciences, e.g., social or economic ones, begin to involve human agency (but see Wagner (2020) for a discussion of typicality in this context), while the boundary between physics and chemistry can be blurry.

15.1 Ontology of Special Sciences

The view of special science laws that I am about to sketch is ontologically reductive. The fundamental ontology of the world is that of fundamental physics. Genes or tigers or economic markets are nothing over and above this physical ontology, but have to be located in it by means of a functionalist analysis. My view, however, is not explanatorily reductive. Biological phenomena are much better explained in biological terms than in terms of atomic trajectories, let's say. Special sciences thus exist not only as a poor substitute for physics, as long as certain systems are too complex for us to provide a complete physical description, but have explanatory autonomy.

To account for this autonomy, I will, in fact, take some hints from Super-Humeanism. The basic idea is that theoretical concepts need not refer directly to (fundamental) properties or entities in the world but may supervene on the regularities as part of their best systematization. While I found this metaphysical strategy untenable when applied to reduce fundamental physical laws and spacetime structure, it strikes me as plausible in the context of special sciences, which deal with non-fundamental laws and entities. In brief, my view is the following:

1. *There are genuine biological regularities in the world.*

   If we think of regularities in terms of complexity theory, that is, roughly in terms of compressible data sets (Kolmogorov complexity), we have already seen that they are highly language-dependent – if we are talking about finite data sets, as we arguably are in the special sciences. In this sense, there are regularities in the world that can only be identified, or at least systematized, in the language of biology (rather than the language of physics), even though they are instantiated in the physical ontology.

While biological terms are in principle translatable into physical terms – by suitable functional definitions or *Ramseyfication* – the functional definition of a gene, or a cell, or a tiger in terms of elementary particles is extremely complex. The translation into the language of physics would thus come at very high costs in terms of simplicity as we try to identify and systematize biological regularities. In addition, there is the issue of *multiple realizability* (see, e.g., Esfeld and Sachse (2007) for a good discussion), that is, one and the same biological term may be realized by different physical states in different instances. Hence, at least in the complexity-theoretic sense, a set of physical events may not instantiate any regular pattern at all (the data set may be "incompressible", or nearly so), unless we introduce appropriate biological terms and macro-variables.

2. *What makes "genes" part of biological laws is their role in the best systematization of biological regularities.*

   For the best system, I adopt (for now) the Mills-Ramsey-Lewis criterion of striking an optimal balance between simplicity and strength. I am open, even sympathetic, to including additional metrics (such as the "cognitive distance" between scientific and manifest image, see Ch. 14) but this is beyond the scope of the present discussion.

   In any case, since we are concerned with the systematization of genuine biological regularities in the sense discussed above, the "nomic status" of the theoretical entity *gene* is provided by biology. This status is what distinguishes genes from spurious or unnatural concepts à la *grue emeralds* that could also be defined in functional physical terms.

   This Humean view corresponds essentially to Loewer's "package deal account" (Loewer, 2007b), according to which the laws and the "natural" properties they are referring to supervene *together* on the regularities as part of their best systematization.

3. However: *Every concrete proposition about genes (and every biological proposition, in general) has a physical truth-maker.*

   Any proposition about the constitution, behavior, or interaction of biological entities is ultimately a proposition about the physical world and thus made true or false by physical facts. In particular, every biological system is also a physical system and must, therefore, obey the laws of physics. No true biological law could ever contradict the physical laws.

   In the upshot, there is nothing over and above the physical facts that makes a biological fact a fact (this is the reductive part of the account). But there is, in

> the sense explained in points 1 and 2, something beyond physics that makes a fact a *biological* fact – with emphasis on the *logos* part (this being the autonomy of the special sciences).

This proposal for the ontological status of non-fundamental entities has certain similarities to Dennett's notion of "real patterns" (Dennett, 1991). I hesitate to adopt it for several reasons, however. First, because the word "pattern" may carry the connotation of something abstract, while I am pretty sure of genes, and damn sure of tigers, that they are concrete entities. Second, my view is quite conservative in that it is reductive, specific to special natural sciences, and (as I will explain below) hierarchical. On the other hand, the main focus of Dennett (1991) are *beliefs* which fall into the mental/normative domain with respect to which I do not advocate for a reductive view, while in more recent literature, "real patterns" are usually tied to radically non-reductionist and non-hierarchical metaphysics ("rainforest realism," see Ladyman and Ross (2007)), or used for functionalist arguments in physics that I consider spurious (see, e.g., Wallace (2003) and our discussion of wave function functionalism in Ch. 12).

Finally, the concept of real patterns has often been invoked in debates about various forms of realism, and I don't feel like I have much to contribute to these debates, nor that they are particularly productive for our present purposes. Genes are real, of course, but they are not fundamental. If we want to call them patterns, then they are patterns instantiated in the physical ontology. (Whether the correct relation between a gene and, let's say, a configuration of elementary particles is one of *identity* or *grounding* is a too subtle metaphysical question for me.) But the same is true of grue emeralds if the term, with a proper functional definition, succeeds in referring at least once. Their different status is due to the fact that genes figure in biological laws, while grue emeralds do not figure in gemological ones.

There is certainly a pragmatic element involved in this distinction (genes are not "more real" than grue emeralds, the concept is just more useful). However, here I would share in the Humean hope that one candidate theory is *objectively* best in its respective domain, whether or not we are able to decide it in practice. I am not positive that this hope is true, but even less convinced of the need to concede to relativism. If two biologists disagree about which part of a DNA sequence is *the gene for blue eyes*, then they either disagree about the meaning of "gene" or (more likely) one of their theories will strike a better balance between simplicity and strength in systematizing blue-eye heredity.

That said, my claims to originality are fairly limited, as my ontological views are drawing a lot from Dennett's real patterns, Loewer's package deals, and the Canberra plan for metaphysics (see Esfeld (2020) for a recent discussion). Notably, though, I am comfortable with using Humean, functionalist, and maybe even pragmatist strategies in my discussion of special sciences because it is rooted in the assumption of a fundamental physical ontology and fundamental (anti-Humean) laws. What I find untenable, if not unintelligible, are real patterns or (Super-)Humeanism "all the way down."

15.2 Probability and Causation in Special Sciences

I now want to relate this discussion of special sciences with our previous analysis of probability and causation in terms of typicality. Indeed, I believe that causal explanations are much more relevant to special sciences than to (fundamental) physics. There is, however, a prima facie tension between the view that biological entities or properties are causally efficacious and the view that they are ontologically reducible to micro-physical entities or properties. What does it mean, for instance, that *a gene mutation increases the fitness of a certain biological form* when the survival and reproductive success of every individual is determined by the physical dynamics guiding its microscopic constituents and the initial conditions of the universe? When there exists, in principle, a complete description of natural evolution in terms of atomic trajectories? It may seem like the biological explanation is either wrong or redundant – unless we had a genuine case of causal overdetermination.

Now, I do not share the latter worry because I don't believe that there are fundamental causal relations in physics. If the dynamical laws are bi-deterministic, then the complete specification of the physical state of the world at one time entails and necessitates its complete state at any other time. But this is not a causal relation, if only for the fact that it is symmetric.

As argued in Ch. 10, causal relations can hold between two macrostates A and B in that one macrostate makes the other typical: $\texttt{Typ}(B \mid A)$ while $\neg\texttt{Typ}(B)$. Such macrostates can be specified in biological (or other special science) terms. Indeed, since we have insisted that biological predicates allow for a translation into the language of physics, they can be conceived as coarse-graining functions on the microscopic state space, i.e., as Boltzmannian macro-variables. We can thus apply our previous physical analysis of probability and causation.

For instance, the fact that an individual has developed the phenotypical trait P makes it typical that S : *it survives long enough to reproduce* in the environmental conditions E.

$$\texttt{Typ}(S \mid E \vee P), \quad \neg\texttt{Typ}(S \mid E) \tag{15.1}$$

Or, if $\{a_1, ..., a_N\}$ is a population with genotype A and $\{b_1, ..., b_M\}$ a population with genotype B (i.e., two statistical ensembles), then the reproduction rate of A may *typically* be greater than the reproduction rate of B. More formally:

$$\texttt{Typ}\left(\frac{1}{N}\sum_{i=1}^{N} S(a_i) > \frac{1}{M}\sum_{j=1}^{M} S(b_j)\right). \tag{15.2}$$

Typicality here is still understood in the usual physical sense of a phenomenon obtaining "for nearly all possible (initial) micro-conditions" (but note the remarks about the context-dependence of typicality in the following section). In the upshot, there is no need for "causal emergentism" or a new source of "randomness." Causal explanations in the special sciences are a form of causal physical inference, as discussed in Ch. 10. And objective probabilities in the special sciences mean the same as physical probabilities, namely typical relative frequencies.

15.3 Special Science Laws as Typicality Laws

Special science laws are generally understood as *ceteris paribus* laws (CP-laws). In contrast to fundamental physical laws, they are not universally true but hold under specific circumstances which exclude interfering factors. The main problem with this concept of (exclusive) CP-laws is that it seems impossible to provide a complete specification of all interfering factors to be excluded (in particular in the language of the respective science) without essentially falling into the tautology that L is true except in circumstances in which it isn't. Hence the charge that CP-laws are in danger of being either false or trivial (see Hempel (1988) and, in particular, Lange (1993)).

It seems to me that this problem arises mostly from attempts to model the concept of special science laws after fundamental physical laws, when they should be really understood as *typical regularities.* Thermodynamic laws would have been a better example to follow if Boltzmann's reduction to statistical mechanics had been more widely appreciated.

To ground explanations, predictions, and counterfactuals, an effective (not fundamental) law need not state conditions that make its instances *necessary.* It only has to be specific enough about its domain – and tolerant enough of small fluctuations – to make the regularities typical. A limited number of definite ceteris paribus clauses will thus belong to the description/systematization of the respective regularity (including the macro-conditions we have to conditionalize on), while the indefinable range of other potential interferences is negligible by virtue of being atypical events. The understanding of special science laws as typical regularities (and in the framework adopted from statistical mechanics) thus leads naturally to a similar conclusion as that expressed by Marc Lange (2002):

> To discover the law that all F's are G, *ceteris paribus*, scientists obviously must understand what factors qualify as 'disturbing'. But they needn't identify all of the factors that can keep an F from being G. They needn't know of factors that, when present, cause only negligible deviations from strict G-hood, or factors that, although capable of causing great departures from G-hood, arise with negligible frequency in the range of cases with which the scientists are concerned. (p. 411)

There are also parallels – as well as important differences – between the typicality view and *normality* theories which understand CP-laws as laws that hold under *normal conditions*, that is, simply put "conditions that normally, usually, mostly obtain" (Spohn, 2008, p. 278). Spohn ultimately rejects this characterization, taking a more epistemic turn to explicate normality conditions in terms of doxastic states and degrees of belief (more precisely, "ranking functions"; see, e.g., Spohn (2002, 2014)). Both definitions of normality differ from typicality. On the one hand, typicality refers first and foremost to what obtains for most possible micro-conditions, i.e., in most nomologically possible worlds, not to what obtains most of the time in the actual world. It is a theorem, rather than a definition, that a repeatable typical event will typically obtain most of the time. On the other hand, what is typical does not depend on anyone's expectations or beliefs. It is the other way round: typicality facts guide rational expectations and beliefs.

15.4 The Hierarchy of Sciences

The theory of a *hierarchy* of modern sciences, often attributed to Auguste Comte (1830), has great intuitive appeal. While the issue can become messy and controversial when one gets into the weeds – and the structure of science is arguably more like a branching tree than a pyramid – it seems by and large correct to say that biological facts reduce to chemical facts, and chemical facts reduce to physical ones. The various levels of this hierarchy are often associated with different scales of size or complexity, different degrees of generality or fundamentality, and sometimes (more judgementally) different degrees of rigor and predictive uncertainty. Here, I want to argue that this hierarchy of sciences is well captured and explained by different "levels" of typicality.

We have already discussed the context-sensitivity of typicality. Simply put, depending on the relevant set of propositions, there may be a different scale of ϵ such that a proposition P is typical if $\mu(P) > 1 - \epsilon$ with respect to a designated typicality measure μ. And it seems plausible that this "threshold" for typicality is very high in the context of physics, lower in the context of chemistry, lower still for propositions relevant to biology, and so on. All propositions are ultimately translatable into physical proposition, leading to a partition of the fundamental microscopic state space into macro-regions. However, which macro-regions are considered "large" or "small" is relative to the particular partition, that is, the context of scientific reasoning.

It is only in a very loose sense that we could associate these "levels of typicality" with (unsharp) degrees of belief. What matters is the rationality principle that we have associated with typicality: A regularity that is typical in the biological context does not require further biological explanation, but may be reducible to typical chemical or physical regularities. And an atypical biological phenomenon requires, in the first place, additional biological explanation (or may compel us to reject our biological theory), but will not, in general, challenge our fundamental theories of physics.

This hierarchy has not just an intuitive appeal but a basis in mathematics if we think of regularities in statistical terms. As we pass to larger and larger scales and more and more specialized sciences, we are dealing with ever-smaller sample sizes. Oversimplified: social regularities are instantiated in systems of $\sim 10^2 - 10^9$ people, physiological regularities are instantiated in systems of $\sim 10^6 - 10^{14}$ cells, and macrophysical regularities are instantiated in systems of $\sim 10^{20} - 10^{80}$ elementary particles. If we think schematically in terms of the law of large numbers

$$\mathbb{P}\left(|\text{relative frequency} - \text{theoretical mean}| > \epsilon\right) = \delta \lesssim \frac{const.}{\epsilon^2 N}, \tag{15.3}$$

we see that smaller sample size N means greater "uncertainty," that is, both a broader range ϵ of typical values, and a larger measure δ of the atypical events. In other words, it is not just an epistemic or methodological issue that biological predictions seem less reliable and precise than, e.g., thermodynamic ones. The fundamental laws of nature and the very scope of the respective regularities make it so.

Remark (Mentaculus account of special science laws)**.** I am not sure if the Humean Mentaculus discussed in Ch. 5 allows for a similar conclusion. Loewer (2012a) provides an account of special science laws based on Humean probabilities for *individual* events, which are understood as the measure of the set of initial micro-conditions of the universe (in the Past Hypothesis macro-region) realizing the respective (macro-)event. In contrast to the typicality theory of probability, there is then no sense in which the Humean chance of a singular political event (let's say) would be less sharp than the Humean chance of a singular physical event. Physicists may be more confident than political scientists, but the Mentaculus has a definite, unwavering opinion about everything. That said, despite our different views about laws and probabilities, I very much concur with Loewer's approach in recognizing Boltzmannian statistical mechanics as the appropriate framework for understanding the emergence of special science laws from physics.

There is another way to understand the hierarchy of sciences, not by reducing each special science directly to physics, but by reducing each higher-level theory to the next lower level.

It is helpful to start with an intra-physical example. We can describe a ball as a rigid Newtonian body with 6 degrees of freedom (3 for the position of its center of mass and 3 rotational degrees of freedom). This is a very coarse-grained description, ignoring the internal (microscopic) degrees of freedom of the object. Technically, we are thereby passing from the microscopic phase space of $\sim 10^{24}$ particles to a reduced or effective 12-dimensional phase space of the ball (6 degrees of freedom and the conjugated momenta), which is good enough to account for most mechanical regularities. In doing so, we are implicitly assuming that each configuration in the reduced phase space is realized by typical micro-states, disregarding micro-conditions for which the ball would suddenly decay, or shoot off an ultra-fast particle while veering in the opposite direction, or perform other shenanigans. Such atypical micro-states would cease to realize a "ball" in the sense relevant to the higher-level theory. On the other hand, if we consider a system of N balls (or "hard spheres") their common effective phase space ($12N$-dimensional) will be equipped with its own natural measure that allows us to make typicality statements with respect to the initial positions and (angular) momenta of the balls.[1]

A similar thing happens as we go from elementary physics to molecular physics, to chemistry, to biology, and so on. In the right combinations and under the right environmental conditions, typical configurations of quarks realize nucleons, typical configurations of nucleons and electrons realize atoms, typical configurations of $\mathrm{O}, \mathrm{N}, \mathrm{H}, \mathrm{C}$ atoms realize cytosine, guanine, adenine or thymine molecules, and typical configurations of those nucleobases (plus deoxyribose and some organic phosphates) realize DNA. In each step, we are passing to a reduced state space, ignoring internal degrees of freedom by assuming typical behavior of the more microscopic constituents.

Admittedly, as we move further away from physics and into more qualitative territory, it becomes questionable whether we still have a meaningful state space, a quantitative conception of a system's degrees of freedom and dynamics that would allow us to formulate precise typicality statements. But taken with a grain of salt, we can say that chemical regularities are instantiated by typical physical systems (more precisely: physical systems of the right kind that behave in a typical manner), biological regularities are instantiated by typical chemical systems (of the right kind), medical regularities are instantiated by typical biological systems (of the right kind), and so on and so forth. In each step, we are multiplying the possibilities of atypical events on the lower levels, which is another way to understand the increase of "uncertainty" – in the sense of a broader range of typical values – and in the size of the "exception sets" of initial micro-conditions on the fundamental micro-physical level.

15.5 Is Life Atypical?

In addition to our previous considerations, it is also interesting to note that many specialized sciences seem to deal only in *conditional* typicality. For instance, economic regularities may be typical *given* the existence of market economies (with more or less rational agents), but to claim that the existence of markets itself is a typical feature of a physical universe seems like a stretch. (Pace Marx, a period of capitalism is not that "inevitable.") On the other hand, physics and maybe chemistry discover also phenomena that are typical *tout court* – which is another way to see why these sciences are more fundamental.

And what about biology? Is the very existence of biological phenomena an atypical feature of our universe? This is a profound and challenging question, not least because it concerns *our* place as intelligent[2] life forms in the cosmos. The main issue, to be clear, is not whether life is ubiquitous in our universe, but whether typical universes allowed by the fundamental laws of nature contain any (complex) life forms at all.

[1] A single ball also has a natural, stationary phase space measure but no interesting statistical regularities.

[2] Although here, I'll be less concerned with the "intelligence" part and bracket the issue of phenomenal consciousness altogether.

I don't have much expertise to offer on the question itself, but by and large, three resolutions seem plausible:

1. The existence of life in the universe is typical because the thermodynamic evolution of the universe – the way in which entropy typically increases – is somehow conducive to the creation of complex subsystems with self-replicating entities that get Darwinian evolution started. (A much-noted exploration of this idea is due to England (2013). Erwin Schrödinger's *What is Life?* (1944) was highly influential in relating the question of life to thermodynamic considerations.)

2. The existence of life is typical – or at least not atypical – merely because of the "large numbers." That is, even though the environmental conditions and the physical processes necessary for the origin of life are extremely special, the universe is so big (and so old) that they are bound to occur *somewhere*.

3. Life is atypical, requiring very fine-tuned initial conditions of the universe.

Let's assume, for the sake of argument, that the last conclusion is correct and the existence of biological life turns out to be atypical according to the fundamental theories of physics. What would the implications be? Wouldn't the fact that our very existence is atypical undermine the rationality of typicality arguments altogether? I believe that, based on typicality reasoning, two different stances could be taken.

The first would simply accept that the phenomenon of life is a challenge to our current best theories. We cannot be content with the fact that life doesn't require a flat-out violation of the physical laws; its atypicality compels us to look additional theoretical principles – or better theories – that make life not atypical. Some may appeal to a (strong) anthropic principle to meet the explanatory burden (see Barrow and Tipler (1986) for a classical reference), but I don't see much value in it.

The other option is more sobering: From the point of view of fundamental physics, the existence of life is an atypical feature of our universe, but it is not a bona fide *physical phenomenon*, i.e., not the kind of phenomenon that physics must be able to explain. It is rather an acceptably brute fact – consistent with the fundamental laws of nature but purely accidental. To me, this option (if correct) would point to much more than a pragmatic division of labor between physics and biology. It would mean that the existence of life – our existence – is insignificant in the great cosmic scheme of things, a footnote in the book of nature.

The idea that our universe is fine-tuned for life is, of course, a popular argument for God, at least a deistic one, who does not have to intervene in the course of nature but set up things very carefully. If the reader will indulge the religious imagery, I would put it the other way around: The question whether God, upon creating the universe, cared about the existence of human beings (more than, let's say, about the exact shape of a certain sand dune changing in the wind), depends on whether God set up laws that make the evolution of intelligent life forms typical.

Bibliography

Albert, D. Z. (2000). *Time and Chance.* Harvard University Press, Cambridge, Massachusetts.

Albert, D. Z. (2013). Wave Function Realism. In Ney, A. and Albert, D. Z., editors, *The Wave Function*, pages 52–57. Oxford University Press, New York.

Albert, D. Z. (2015). *After Physics.* Harvard University Press, Cambridge, Massachusetts.

Allori, V., Goldstein, S., Tumulka, R., and Zanghì, N. (2008). On the common structure of Bohmian mechanics and the Ghirardi-Rimini-Weber theory. *British Journal for the Philosophy of Science*, 59(3):353–389.

Allori, V., Goldstein, S., Tumulka, R., and Zanghì, N. (2014). Predictions and primitive ontology in quantum foundations: A study of examples. *British Journal for the Philosophy of Science*, 65(2):323–352.

Armstrong, D. M. (1983). *What Is a Law of Nature?* Cambridge University Press, Cambridge.

Armstrong, D. M. (1986). The Nature of Possibility. *Canadian Journal of Philosophy*, 16(4):575–594.

Armstrong, D. M. (1989). *A Combinatorial Theory of Possibility.* Cambridge University Press, Cambridge.

Arnold, V. I. and Avez, A. (1968). *Ergodic Problems of Classical Mechanics.* The Mathematical Physics Monograph Series. W.A. Benjamin, first edition edition.

Ash, R. B. and Doleans-Dade, C. A. (2000). *Probability and Measure Theory.* Academic Press.

Bacciagaluppi, G. and Valentini, A. (2009). *Quantum Theory at the Crossroads: Reconsidering the 1927 Solvay Conference.* Cambridge University Press, Cambridge.

Barbour, J. (2003). Scale-invariant gravity: Particle dynamics. *Classical and Quantum Gravity*, 20:1543–1570.

Barbour, J. (2017). Arrows of Time in Unconfined Systems. In Renner, R. and Stupar, S., editors, *Time in Physics*, Tutorials, Schools, and Workshops in the Mathematical Sciences, pages 17–26. Springer International Publishing, Cham.

Barbour, J. and Bertotti, B. (1982). Mach's principle and the structure of dynamical theories. *Proceedings of the Royal Society A*, 382:295–306.

Barbour, J., Koslowski, T., and Mercati, F. (2013). A Gravitational Origin of the Arrows of Time. *arXiv:1310.5167 [astro-ph, physics:gr-qc]*.

Barbour, J., Koslowski, T., and Mercati, F. (2014). Identification of a Gravitational Arrow of Time. *Physical Review Letters*, 113(18):181101.

Barbour, J., Koslowski, T., and Mercati, F. (2015). Entropy and the Typicality of Universes. *arXiv:1507.06498 [gr-qc]*.

Barrett, J. A. (2001). *The Quantum Mechanics of Minds and Worlds*. Oxford University Press, Oxford, New York, first edition.

Barrett, J. A. (2016). Typicality in Pure Wave Mechanics. *Fluctuation and Noise Letters*, 15(03):1640009.

Barrow, J. D. and Tipler, F. J. (1986). *The Anthropic Cosmological Principle*. Oxford University Press, Oxford; New York.

Bell, J. S. (2004). *Speakable and Unspeakable in Quantum Mechanics*. Cambridge: Cambridge University Press, second edition.

Belot, G. (2000). Geometry and motion. *British Journal for the Philosophy of Science*, 51(4):561–595.

Berndl, K., Dürr, D., Goldstein, S., Peruzzi, G., and Zanghì, N. (1995). On the global existence of Bohmian mechanics. *Communications in Mathematical Physics*, 173(3):647–673.

Bernoulli, J. (1713). *Ars Conjectandi*. Impensis Thurnisiorum, Fratrum.

Bhogal, H. and Perry, Z. R. (2017). What the Humean should say about entanglement. *Noûs*, 51(1):DOI 10.1111/nous.12095.

Bird, A. (2007). *Nature's Metaphysics: Laws and Properties*. New York: Oxford University Press.

Birkhoff, G. D. (1931). Proof of the ergodic theorem. *Proceedings of the National Academy of Sciences of the United States of America*, 17:656–660.

Bohm, D. (1952a). A suggested interpretation of the quantum theory in terms of "hidden" variables. 1. *Physical Review*, 85(2):166–179.

Bohm, D. (1952b). A suggested interpretation of the quantum theory in terms of "hidden" variables. 2. *Physical Review*, 85(2):180–193.

Boltzmann, L. (1896a). Entgegnung auf die wärmetheoretischen Betrachtungen des Hrn. E. Zermelo. *Wiedemanns Annalen*, 57:773–784.

Boltzmann, L. (1896b). *Vorlesungen über Gastheorie*, volume 1. J. A. Barth, Leipzig.

Boltzmann, L. (1897). Zu Hrn. Zermelos Abhandlung 'Über die mechanische Erklärung irreversibler Vorgänge'. *Annalen der Physik*, 60:392–398.

Borel, E. (1909). Les probabilités dénombrables et leurs applications arithmétiques. *Rendiconti del Circolo Matematico di Palermo (1884-1940)*, 27(1):247–271.

Borel, E. (1939). *Valeur pratique et philosophie des probabilités.* Gauthier-Villars, Paris.

Borel, É. (1948). *Le hasard.* Presses universitaires de France, Paris.

Bricmont, J. (1995). Science of Chaos or Chaos in Science? *Annals of the New York Academy of Sciences*, 775(1):131–175.

Bricmont, J. (2001). Bayes, Boltzmann and Bohm: Probabilities in Physics. In *Chance in Physics*, Lecture Notes in Physics, pages 3–21. Springer, Berlin, Heidelberg.

Bricmont, J. (2016). *Making Sense of Quantum Mechanics.* Springer International Publishing, Cham.

Brush, S. G. (1966). *Kinetic Theory: Irreversible Processes*, volume 2. Elsevier.

Bunimovich, L. A. (1979). On the ergodic properties of nowhere dispersing billiards. *Communications in Mathematical Physics*, 65(3):295–312.

Callender, C. (2004a). Measures, Explanations and the Past: Should 'Special' Initial Conditions be Explained? *The British Journal for the Philosophy of Science*, 55(2):195–217.

Callender, C. (2004b). There is No Puzzle About the Low-Entropy Past. In Hitchcock, C., editor, *Contemporary Debates in Philosophy of Science*, pages 240–255. Blackwell.

Callender, C. (2007). The emergence and interpretation of probability in Bohmian mechanics. *Studies in History and Philosophy of Modern Physics*, 38:351–370.

Callender, C. (2010). The past hypothesis meets gravity. In Ernst, G. and Hüttemann, A., editors, *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics*, pages 34–58. Cambridge University Press, Cambridge.

Callender, C. (2015). One world, one beable. *Synthese*, 192(10):3153–3177.

Callender, C. (2016). Thermodynamic Asymmetry in Time. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition.

Carroll, J. W. (1994). *Laws of Nature*. Cambridge University Press, Cambridge.

Carroll, S. (2010). *From Eternity to Here*. Dutton, New York.

Carroll, S. M. and Chen, J. (2004). Spontaneous Inflation and the Origin of the Arrow of Time. *arXiv:hep-th/0410270*.

Carroll, S. M. and Tam, H. (2010). Unitary Evolution and Cosmological Fine-Tuning.

Cartwright, N. (1983). *How the Laws of Physics Lie*. Oxford University Press, Oxford, New York.

Chaitin, G. J. (2007). *Thinking about Godel and Turing: Essays on Complexity, 1970-2007*. World Scientific.

Chalmers, D. J. (1995). Facing Up to the Problem of Consciousness. *Journal of Consciousness Studies*, 2(3):200–19.

Chen, E. K. (2018). Quantum Mechanics in a Time-Asymmetric Universe: On the Nature of the Initial Quantum State. *The British Journal for the Philosophy of Science*.

Cohen, J. and Callender, C. (2009). A better best system account of lawhood. *Philosophical Studies*, 145(1):1–34.

Cohen-Tannoudji, C., Diu, B., and Laloe, F. (1991). *Quantum Mechanics, Vol. 1*. Wiley, New York, 1st edition.

Coles, P. and Ellis, G. (1997). *Is the Universe Open or Closed?: The Density of Matter in the Universe*. Cambridge Lecture Notes in Physics. Cambridge University Press, Cambridge, first edition.

Comte, A. (1830). *Cours de philosophie positive*. Bachelier.

Cournot, A. A. (1843). *Exposition de la théorie des chances et des probabilités*. L. Hachette.

Cowan, C. W. and Tumulka, R. (2016). Epistemology of wave function collapse in quantum physics. *British Journal for the Philosophy of Science*, 67:405–434.

Crane, H. and Wilhelm, I. (2020). The Logic of Typicality. In Allori, V., editor, *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism, and Laws of Nature*. World Scientific.

Curiel, E. (2015). Measure, Topology and Probabilistic Reasoning in Cosmology. *arXiv:1509.01878 [gr-qc, physics:math-ph, physics:physics].*

Das, S. and Dürr, D. (2019). Arrival Time Distributions of Spin-1/2 Particles. *Scientific Reports*, 9(1):1–8.

Daumer, M., Dürr, D., Goldstein, S., and Zanghì, N. (1996). Naive realism about operators. *Erkenntnis*, 45(2):379–397.

Davies, P. C. W. (1977). *The Physics of Time Asymmetry.* University of California Press, Berkeley and Los Angeles.

de Broglie, L. (1928). La nouvelle dynamique des quanta. In *Electrons et Photons: Rapports et Discussions Du Cinquième Conseil de Physique.*, pages 105–132. Gauthier-Villars, Paris.

Dennett, D. C. (1991). Real Patterns. *Journal of Philosophy*, 88(1):27–51.

Dretske, F. I. (1977). Laws of Nature. *Philosophy of Science*, 44(2):248–268.

Duffin, R. J. and Schaeffer, A. C. (1941). Khintchine's problem in metric Diophantine approximation. *Duke Mathematical Journal*, 8(2):243–255.

Duhem, P. M. M. (1954). *The Aim and Structure of Physical Theory.* Princeton University Press, Princeton.

Dürr, D., Froemel, A., and Kolb, M. (2017). *Einführung in Die Wahrscheinlichkeitstheorie Als Theorie Der Typizität.* Springer Berlin Heidelberg, Berlin, Heidelberg.

Dürr, D., Goldstein, S., Norsen, T., Struyve, W., and Zanghì, N. (2013a). Can Bohmian mechanics be made relativistic? *Proceedings of the Royal Society A*, 470:2162.

Dürr, D., Goldstein, S., and Zanghì, N. (1992). Quantum equilibrium and the origin of absolute uncertainty. *Journal of Statistical Physics*, 67(5-6):843–907.

Dürr, D., Goldstein, S., and Zanghì, N. (1997). Bohmian Mechanics and the Meaning of the Wave Function. In Cohen, R. S., Horne, M., and Stachel, J. J., editors, *Experimental Metaphysics: Quantum Mechanical Studies for Abner Shimony, Volume One*, Boston Studies in the Philosophy and History of Science, pages 25–38. Springer Netherlands.

Dürr, D., Goldstein, S., and Zanghì, N. (2004). Quantum Equilibrium and the Role of Operators as Observables in Quantum Theory. *Journal of Statistical Physics*, 116(1):959–1055.

Dürr, D., Goldstein, S., and Zanghì, N. (2013b). *Quantum Physics without Quantum Philosophy.* Berlin: Springer.

Dürr, D., Goldstein, S., and Zanghì, N. (2019). Quantum Motion on Shape Space and the Gauge Dependent Emergence of Dynamics and Probability in Absolute Space and Time. *Journal of Statistical Physics*.

Dürr, D. and Lazarovici, D. (2018). *Verständliche Quantenmechanik: Drei Mögliche Weltbilder Der Quantenphysik*. Springer Spektrum.

Dürr, D. and Lazarovici, D. (2020). *Understanding Quantum Mechanics : The World According to Modern Quantum Foundations*. Springer International Publishing.

Dürr, D. and Teufel, S. (2009). *Bohmian Mechanics: The Physics and Mathematics of Quantum Theory*. Springer, Berlin.

Earman, J. (1986). *A Primer on Determinism*. The Western Ontario Series in Philosophy of Science. Springer Netherlands.

Ehrenfest, P. a. T. (1907). Begriffliche Grundlagen der Statistischen Auffassung in der Mechanik. In Klein, F. and Müller, C., editors, *Mechanik*, pages 773–860. Vieweg+Teubner Verlag, Wiesbaden.

Einstein, A. (1948). Quanten-Mechanik und Wirklichkeit. *Dialectica*, 2:320–324.

Einstein, A. (1987). *Letters to Solovine*. Philosophical Library.

Emery, N. (2019). Laws and their instances. *Philosophical Studies*, 176(6):1535–1561.

England, J. L. (2013). Statistical physics of self-replication. *The Journal of Chemical Physics*, 139(12):121923.

Erdős, P. and Rényi, A. (1963). Asymmetric graphs. *Acta Mathematica Academiae Scientiarum Hungarica*, 14(3):295–315.

Esfeld, M. (2014a). The primitive ontology of quantum physics: Guidelines for an assessment of the proposals. *Studies in History and Philosophy of Modern Physics*, 47:99–106.

Esfeld, M. (2014b). Quantum Humeanism, or: Physicalism without properties. *The Philosophical Quarterly*, 64(256):453–470.

Esfeld, M. (2018). Collapse or No Collapse? What Is the Best Ontology of Quantum Mechanics in the Primitive Ontology Framework? In Gao, S., editor, *Collapse of the Wave Function: Models, Ontology, Origin, and Implications*, pages 167–184. Cambridge University Press, Cambridge.

Esfeld, M. (2020). Super-Humeanism: The Canberra Plan for Physics. In Glick, D., Darby, G., and Marmodoro, A., editors, *The Foundation of Reality: Fundamentality, Space, and Time*, page Chapter 6. Oxford University Press, Oxford, New York.

Esfeld, M. and Deckert, D.-A. (2017). *A Minimalist Ontology of the Natural World*. Routledge Studies in the Philosophy of Mathematics and Physics. Routledge, Oxford.

Esfeld, M., Lazarovici, D., Hubert, M., and Dürr, D. (2014). The ontology of Bohmian mechanics. *British Journal for the Philosophy of Science*, 65(4):773–796.

Esfeld, M., Lazarovici, D., Lam, V., and Hubert, M. (2015). The Physics and Metaphysics of Primitive Stuff. *The British Journal for the Philosophy of Science*, 68(1):133–161.

Esfeld, M. and Sachse, C. (2007). Theory Reduction by Means of Functional Sub-types. *International Studies in the Philosophy of Science*, 21(1):1–17.

Everett, H. (1956). The Theory of the Universal Wave Function.

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

Everett, H. (1973). The Theory of the Universal Wave Function. *The Many-Worlds Interpretation of Quantum Mechanics*, pages 3–140.

Farr, M. (2018). Causation and Time Reversal. *The British Journal for the Philosophy of Science*.

Feynman, R. (1967). *The Character of Physical Law*. M.I.T. Press.

Finkelberg, A. (2017). *Heraclitus and Thales' Conceptual Scheme: A Historical Study*. BRILL.

Fletcher, S. C. (2020). The Principle of Stability. *Philosophers' Imprint*, 20(3).

Foster, J. (2004). *The Divine Lawmaker*. Oxford University Press.

Frigg, R. (2009). Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics. *Philosophy of Science*, 76(5):997–1008.

Frigg, R. (2011). Why Typicality Does Not Explain the Approach to Equilibrium. In Suárez, M., editor, *Probabilities, Causes and Propensities in Physics*, Synthese Library, pages 77–93. Springer Netherlands, Dordrecht.

Frigg, R. and Werndl, C. (2011). Explaining Thermodynamic-Like Behavior in Terms of Epsilon-Ergodicity. *Philosophy of Science*, 78(4):628–652.

Frigg, R. and Werndl, C. (2012). Demystifying Typicality. *Philosophy of Science*, 79(5):917–929.

Georgii, H.-O. (2004). *Stochastik: Einführung in die Wahrscheinlichkeitstheorie und Statistik*. De Gruyter, Berlin; New York, second edition.

Ghirardi, G. C., Grassi, R., and Benatti, F. (1995). Describing the macroscopic world: Closing the circle within the dynamical reduction program. *Foundations of Physics*, 25(1):5–38.

Ghirardi, G. C., Rimini, A., and Weber, T. (1986). Unified dynamics for microscopic and macroscopic systems. *Physical Review D*, 34(2):470–491.

Goldstein, S. (2001). Boltzmann's Approach to Statistical Mechanics. In Bricmont, J., Dürr, D., Galavotti, M. C., Ghirardi, G., Petruccione, F., and Zanghì, N., editors, *Chance in Physics: Foundations and Perspectives*, pages 39–54. Springer, Berlin.

Goldstein, S. (2012). Typicality and Notions of Probability in Physics. In *Probability in Physics*, The Frontiers Collection, pages 59–71. Springer, Berlin, Heidelberg.

Goldstein, S. (2019). Individualist and Ensemblist Approaches to the Foundations of Statistical Mechanics. *The Monist*, 102(4):439–457.

Goldstein, S. and Lebowitz, J. L. (2004). On the (Boltzmann) entropy of non-equilibrium systems. *Physica D: Nonlinear Phenomena*, 193(1):53–66.

Goldstein, S., Lebowitz, J. L., Mastrodonato, C., Tumulka, R., and Zanghì, N. (2010). Approach to thermal equilibrium of macroscopic quantum systems. *Physical Review E*, 81(1):011109.

Goldstein, S., Lebowitz, J. L., Tumulka, R., and Zanghì, N. (2017). Any orthonormal basis in high dimension is uniformly distributed over the sphere. *Annales de l'Institut Henri Poincaré, Probabilités et Statistiques*, 53(2):701–717.

Goldstein, S. and Struyve, W. (2007). On the uniqueness of quantum equilibrium in Bohmian mechanics. *Journal of Statistical Physics*, 128(5):1197–1209.

Goldstein, S., Tumulka, R., and Zanghì, N. (2016). Is the hypothesis about a low entropy initial state of the Universe necessary for explaining the arrow of time? *Physical Review D*, 94(2):023520.

Goldstein, S. and Zanghì, N. (2013). Reality and the Role of the Wave Function in Quantum Theory. In Dürr, D., Goldstein, S., and Zanghì, N., editors, *Quantum Physics Without Quantum Philosophy*, pages 263–278. Springer, Berlin, Heidelberg.

Hacking, I. (1975). The identity of indiscernibles. *Journal of Philosophy*, 72:249–256.

Hadamard, J. (1902). Sur les problèmes aux dérivées partielles et leur signification physique. *Princeton University Bulletin*, 13:49–52.

Hájek, A. (1996). "Mises redux" — Redux: Fifteen arguments against finite frequentism. *Erkenntnis*, 45(2):209–227.

Hájek, A. (2006). The reference class problem is your problem too. *Synthese*, 156:563–585.

Hájek, A. (2009). Fifteen Arguments against Hypothetical Frequentism. *Erkenntnis (1975-)*, 70(2):211–235.

Hall, N. (1994). Correcting The Guide to Objective Chance. *Mind*, 103(412):505–518.

Hall, N. (2004). Two Mistakes About Credence and Chance. *Australasian Journal of Philosophy*, 82(1):93–111.

Hall, N. (2009). Humean reductionism about laws of nature.

Hauray, M. and Jabin, P.-E. (2015). Particle approximation of Vlasov equations with singular forces: Propagation of chaos. *Annales scientifiques de l'École normale supérieure*, 48(4):891–940.

Hawking, S. W. and Ellis, G. F. R. (1973). *The Large Scale Structure of Space-Time.* Cambridge University Press.

Hawking, S. W. and Page, D. N. (1988). How probable is inflation? *Nuclear Physics B*, 298(4):789–809.

Heggie, D. and Hut, P. (2003). *The Gravitational Million-Body Problem.* Cambridge University Press, Cambridge.

Helbig, P. (2012). Is there a flatness problem in classical cosmology? *Monthly Notices of the Royal Astronomical Society.*

Hempel, C. G. (1988). Provisoes: A problem concerning the inferential function of scientific theories. *Erkenntnis*, 28(2):147–164.

Hildebrand, T. (2013). Can Primitive Laws Explain? *Philosopher's Imprint*, 13(15).

Hoefer, C. (2002). Freedom from the Inside Out. *Royal Institute of Philosophy Supplements*, 50:201–222.

Hubert, M. (2019). Typicality and Atypicality: Unifying Probabilities and Really Statistical Explanations.

Huggett, N. (2006). The regularity account of relational spacetime. *Mind*, 115(457):41–73.

Jackson, F. (1994). Armchair metaphysics. In Michael, M. and Hawthorne, J. O., editors, *Philosophy in Mind. The Place of Philosophy in the Study of Mind*, pages 23–42. Dordrecht: Kluwer.

Jeans, J. H. (1915). On the Theory of Star-Streaming and the Structure of the Universe. *Monthly Notices of the Royal Astronomical Society*, 76(2):70–84.

Kac, M. (1959). *Statistical Independence in Probability, Analaysis, and Number Theory.* The Carus Mathematical Monographs. Mathematical Association of America, Washington DC.

Khinchin, A. (1926). Zur metrischen Theorie der diophantischen Approximationen. *Mathematische Zeitschrift*, 24(1):706–714.

Kiefer, C. (2015). Does time exist in quantum gravity? *Philosophical Problems in Science (Zagadnienia Filozoficzne w Nauce)*, 59(59):7–24.

Kiessling, M. K.-H. (2001). How to Implement Boltzmann's Probabilistic Ideas in a Relativistic World? In Bricmont, J., Ghirardi, G., Dürr, D., Petruccione, F., Galavotti, M. C., and Zanghi, N., editors, *Chance in Physics: Foundations and Perspectives*, Lecture Notes in Physics, pages 83–100. Springer, Berlin, Heidelberg.

Kim, J. (1986). Possible Worlds and Armstrong's Combinatorialism. *Canadian Journal of Philosophy*, 16(4):595–612.

Kolomogoroff, A. (1933). *Grundbegriffe Der Wahrscheinlichkeitsrechnung.* Ergebnisse Der Mathematik Und Ihrer Grenzgebiete. 1. Folge. Springer-Verlag, Berlin Heidelberg.

Koukoulopoulos, D. and Maynard, J. (2019). On the Duffin-Schaeffer conjecture. *arXiv:1907.04593 [math].*

Kyburg, H. E. (1961). *Probability and the Logic of Rational Belief.* Wesleyan University Press.

Ladyman, J. and Ross, D. (2007). *Every Thing Must Go: Metaphysics Naturalized.* Oxford University Press, Oxford.

Lake, K. (2005). The Flatness Problem and $\ensuremath{\Lambda}$. *Physical Review Letters*, 94(20):201102.

Lanford, O. E. I. (1975). Time evolution of large classical systems. In Moser, J., editor, *Dynamical Systems, Theory and Applications*, volume 38 of *Lecture Notes in Physics*, pages 1–111. Springer, Berlin, Heidelberg.

Lange, M. (1993). Natural laws and the problem of provisos. *Erkenntnis*, 38(2):233–248.

Lange, M. (2002). Who's Afraid of Ceteris-Paribus Laws? Or: How I Learned to Stop Worrying and Love Them. *Erkenntnis*, 57(3):407–423.

Lange, M. (2009). *Laws and Lawmakers: Science, Metaphysics, and the Laws of Nature.* Oxford University Press, Oxford, New York.

Lange, M. (2013). Grounding, scientific explanation, and Humean laws. *Philosophical Studies*, 164:255–261.

Lavis, D. A. (2005). Boltzmann and Gibbs: An attempted reconciliation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 36(2):245–273.

Lazarovici, D. (2018). Against fields. *European Journal for Philosophy of Science*, 8(2):145–170.

Lazarovici, D. and Hubert, M. (2019). How Quantum Mechanics can consistently describe the use of itself. *Scientific Reports*, 9(1):470.

Lazarovici, D., Oldofredi, A., and Esfeld, M. (2018). Observables and Unobservables in Quantum Mechanics: How the No-Hidden-Variables Theorems Support the Bohmian Particle Ontology. *Entropy*, 20(5):381.

Lazarovici, D. and Pickl, P. (2017). A Mean Field Limit for the Vlasov–Poisson System. *Archive for Rational Mechanics and Analysis*, 225(3):1201–1231.

Lazarovici, D. and Reichert, P. (2015). Typicality, Irreversibility and the Status of Macroscopic Laws. *Erkenntnis*, 80(4):689–716.

Lebowitz, J. L. (1993a). Boltzmann's Entropy and Time's Arrow. *Physics Today*, 46(9, 32).

Lebowitz, J. L. (1993b). Macroscopic laws, microscopic dynamics, time's arrow and Boltzmann's entropy. *Physica A*, 194(1-4).

Leibniz, G. W. (1982). *New Essays on Human Understanding.* Cambridge University Press, Cambridge, abridged edition. Original work published in 1765.

Leitgeb, H. (2014). The Stability Theory of Belief. *Philosophical Review*, 123(2):131–171.

Leitgeb, H. (2017). *The Stability of Belief: How Rational Belief Coheres with Probability.* Oxford University Press, Oxford.

Lewis, D. (1973). Causation. *Journal of Philosophy*, 70:556–567.

Lewis, D. (1979). Counterfactual Dependence and Time's Arrow. *Noûs*, 13(4):455–476.

Lewis, D. (1980). A Subjectivist's Guide to Objective Chance. In Harper, W. L., Stalnaker, R., and Pearce, G., editors, *IFS: Conditionals, Belief, Decision, Chance and Time*, The University of Western Ontario Series in Philosophy of Science, pages 267–297. Springer Netherlands, Dordrecht.

Lewis, D. (1983). New work for a theory of universals. *Australasian Journal of Philosophy*, 61(4):343–377.

Lewis, D. (1986a). *On the Plurality of Worlds.* Blackwell, Oxford.

Lewis, D. (1986b). *Philosophical Papers*, volume 2. Oxford University Press, Oxford.

Lewis, D. (1994). Humean supervenience debugged. *Mind*, 103(412):473–490.

Lewis, G. F. and Barnes, L. A. (2016). *A Fortunate Universe.* Cambridge University Press.

Lewis, P. J. (2007). How Bohm's Theory Solves the Measurement Problem. *Philosophy of Science*, 74(5):749–760.

Loewer, B. (1996). Humean Supervenience. *Philosophical Topics*, 24:101–127.

Loewer, B. (2001). Determinism and Chance. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 32(4):609–620.

Loewer, B. (2004). David Lewis's Humean Theory of Objective Chance. *Philosophy of Science*, 71(5):1115–1125.

Loewer, B. (2007a). Counterfactuals and the Second Law. In Price, H. and Corry, R., editors, *Causation, Physics, and the Constitution of Reality. Russell's Republic Revisited*, pages 293–326. Oxford University Press, Oxford.

Loewer, B. (2007b). Laws and Natural Properties. *Philosophical Topics*, 35(1/2):313–328.

Loewer, B. (2012a). The emergence of time's arrows and special science laws from physics. *Interface Focus*, 2(1):13–19.

Loewer, B. (2012b). Two accounts of laws and time. *Philosophical Studies*, 160(1):115–137.

Loewer, B. (2020). The Mentaculus Vision. In Allori, V., editor, *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism, and Laws of Nature.* World Scientific.

Loewer, B. (forthcoming). The Consequence Argument Meets the Mentaculus.

López, C. (2019). Roads to the past: How to go and not to go backward in time in quantum theories. *European Journal for Philosophy of Science*, 9(2):27.

Mack, K. (2015). Death in a vacuum. *Cosmos*, 64(Aug-Sep 2015).

Marchal, C. and Saari, D. G. (1976). On the final evolution of the n-body problem. *Journal of Differential Equations*, 20(1):150–186.

Markov, A. A. (1912). *Wahrscheinlichkeitsrechnung.* B.G. Teubner, Leipzig, Berlin.

Martin, T. (1996). *Probabilités et critique philosophique selon Cournot.* Librairie Philosophique J. VRIN.

Maudlin, T. (1995). Three measurement problems. *Topoi*, 14:7–15.

Maudlin, T. (1997). Descrying the World in the Wave Function. *The Monist*, 80(1):3–23.

Maudlin, T. (2007a). *The Metaphysics Within Physics.* Oxford University Press, Oxford.

Maudlin, T. (2007b). What could be objective about probabilities? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 38(2):275–291.

Maudlin, T. (2010). Can the world be only wave-function? In Saunders, S., Barrett, J., Kent, A., and Wallace, D., editors, *Many Worlds? Everett, Quantum Theory, and Reality*, pages 121–143. Oxford University Press, Oxford.

Maudlin, T. (2011). *Quantum Non-Locality and Relativity. Third Edition.* Wiley-Blackwell.

Maudlin, T. (2014). Critical Study David Wallace, The Emergent Multiverse: Quantum Theory According to the Everett Interpretation. Oxford University Press, 2012, 530 + xv pp. *Noûs*, 48(4):794–808.

Maudlin, T. (2019). *Philosophy of Physics: Quantum Theory.* Princeton University Press, Princeton.

Maudlin, T. (2020). The Grammar of Typicality. In Allori, V., editor, *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism and Laws of Nature.* World Scientific.

McCoy, C. D. (2017). Can Typicality Arguments Dissolve Cosmology's Flatness Problem? *Philosophy of Science*, 84(5):1239–1252.

Miller, E. (2014). Quantum entanglement, Bohmian mechanics, and Humean supervenience. *Australasian Journal of Philosophy*, 92:567–583.

Monton, B. (2006). Quantum mechanics and 3N-dimensional space. *Philosophy of science*, 73(5):778–789.

Myrvold, W. C. (2012). Deterministic Laws and Epistemic Chances. In Ben-Menahem, Y. and Hemmo, M., editors, *Probability in Physics*, The Frontiers Collection, pages 73–85. Springer, Berlin, Heidelberg.

Myrvold, W. C. (2016). Probabilities in Statistical Mechanics. In Hitchcock, C. and Hájek, A., editors, *The Oxford Handbook of Probability and Philosophy*, pages 573–600. Oxford University Press.

Myrvold, W. C. (2020). Explaining Thermodynamics: What remains to be done? In Allori, V., editor, *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism, and Laws of Nature*. World Scientific.

Ney, A. (2015). Fundamental physical ontologies and the constraint of empirical coherence: A defense of wave function realism. *Synthese*, 192(10):3105–3124.

Norsen, T. (2017). *Foundations of Quantum Mechanics: An Exploration of the Physical Meaning of Quantum Theory*. Undergraduate Lecture Notes in Physics. Springer International Publishing, Cham.

Norton, J. D. (2008). The Dome: An Unexpectedly Simple Failure of Determinism. *Philosophy of Science*, 75(5):786–798.

Padmanabhan, T. (1990). Statistical mechanics of gravitating systems. *Physics Reports*, 188(5):285–362.

Painlevé, P. (1897). *Leçons Sur La Théorie Analytique Des Équations Différentielles*. A. Hermann, Paris.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics*. Oxford University Press, Oxford.

Pollard, H. (1967). The Behavior of Gravitational Systems. *Journal of Mathematics and Mechanics*, 17(6):601–611.

Price, H. (1996). *Time's Arrow and Archimedes' Point. New Directions for the Physics of Time*. Oxford University Press, Oxford.

Price, H. (2002). Burbury's Last Case: The Mystery of the Entropic Arrow. In Callender, C., editor, *Time, Reality & Experience*, pages 19–56. Cambridge University Press, Cambridge.

Putnam, H. (1969). Is Logic Empirical? In Cohen, R. S. and Wartofsky, M. W., editors, *Boston Studies in the Philosophy of Science: Proceedings of the Boston Colloquium for the Philosophy of Science 1966/1968*, Boston Studies in the Philosophy of Science, pages 216–241. Springer Netherlands, Dordrecht.

Quine, W. V. (1951). Main Trends in Recent Philosophy: Two Dogmas of Empiricism. *The Philosophical Review*, 60(1):20–43.

Reichenbach, H. (1956). *The Direction of Time*. Dover Publications, Mineola, New York.

Reichert, P. (2012). *Can a Parabolic-like Evolution of the Entropy of the Universe Provide the Foundation for the Second Law of Thermodynamics?* Master Thesis, LMU, Munich.

Reichert, P. (2020). Essentially Ergodic Behaviour. *The British Journal for the Philosophy of Science.*

Russell, B. (1912). On the notion of cause. *Proceedings of the Aristotelian Society*, 13:1–26.

Saari, D. G. (1971a). Expanding gravitational systems. *Transactions of the American Mathematical Society*, 156:219–240.

Saari, D. G. (1971b). Improbability of Collisions in Newtonian Gravitational Systems. *Transactions of the American Mathematical Society*, 162:267–271.

Saari, D. G. (1973). Improbability of Collisions in Newtonian Gravitational Systems. II. *Transactions of the American Mathematical Society*, 181:351–368.

Saari, D. G. (2005). *Collisions, Rings, and Other Newtonian N-Body Problems.* Number Nr. 104 in CBMS Regional Conference Series in Mathematics. American Mathematical Society.

Saunders, S. (2013). Rethinking Newton's Principia. *Philosophy of Science*, 80(1):22–48.

Schaffer, J. (2008). Causation and Laws of Nature : Reductionism. In Sider, T., Hawthorne, J., and Zimmerman, D. W., editors, *Contemporary Debates in Metaphysics*, pages 82–107. Blackwell.

Schaffer, J. (2009). Spacetime the one substance. *Philosophical Studies*, 145(1):131–148.

Schaffer, J. (2016). It is the Business of Laws to Govern. *Dialectica*, 70(4):577–588.

Schiffrin, J. S. and Wald, R. M. (2012). Measure and probability in cosmology. *Physical Review D*, 86(2):023521.

Schilpp, P., editor (1949). *Albert Einstein: Philosopher-Scientist.* Number VII in The Library of Living Philosophers. The Library of Living Philosophers Inc., Evanston, Illinois, 1st edition.

Schrodinger, E. (1944). *What Is Life?* Cambridge University Press, Cambridge, reprint edition.

Schwartz, J. (1966). The Pernicious Influence of Mathematics On Science. In Nagel, E., Suppes, P., and Tarski, A., editors, *Studies in Logic and the Foundations of Mathematics*, volume 44 of *Logic, Methodology and Philosophy of Science*, pages 356–360. Elsevier.

Schwarz, W. (2014). Proving the Principal Principle. In Wilson, A., editor, *Chance and Temporal Asymmetry*, pages 81–99. Oxford University Press.

Sebens, C. T. and Carroll, S. M. (2016). Self-locating Uncertainty and the Origin of Probability in Everettian Quantum Mechanics. *The British Journal for the Philosophy of Science*, 69(1):25–74.

Sellars, W. (1962). Philosophy and the scientific image of man. In Colodny, R., editor, *Frontiers of Science and Philosophy*, pages 35–78. University of Pittsburgh Press, Pittsburgh.

Shafer, G. (1985). Conditional Probability. *International Statistical Review / Revue Internationale de Statistique*, 53(3):261–275.

Shafer, G. and Vovk, V. (2006). The Sources of Kolmogorov's Grundbegriffe. *Statistical Science*, 21(1):70–98.

Sider, T. (2005). Another Look at Armstrong's Combinatorialism. *Noûs*, 39(4):679–695.

Silver, N. (2012). *The Signal and the Noise: Why So Many Predictions Fail – but Some Don't.* Penguin, New York.

Sinai, Y. G. (1970). Dynamical systems with elastic reflections. *Russian Mathematical Surveys*, 25(2):137.

Sklar, L. (1973). Statistical Explanation and Ergodic Theory. *Philosophy of Science*, 40(2):194–212.

Spohn, H. (1991). *Large Scale Dynamics of Interacting Particles.* Springer, Berlin, Heidelberg.

Spohn, W. (2002). Laws, Ceteris Paribus Conditions, and the Dynamics of Belief. *Erkenntnis*, 57(3):373–394.

Spohn, W. (2008). *Causation, Coherence and Concepts: A Collection of Essays.* Springer Science & Business Media.

Spohn, W. (2014). The epistemic account of ceteris paribus conditions. *European Journal for Philosophy of Science*, 4(3):385–408.

Streater, R. F. and Wightman, A. S. (2000). *PCT, Spin and Statistics, and All That.* Princeton University Press, Princeton, N.J, revised edition.

Taleb, N. N. (2010). *The Black Swan: The Impact of the Highly Improbable.* Random House Publishing Group, second edition.

Teufel, S. and Tumulka, R. (2005). Simple Proof for Global Existence of Bohmian Trajectories. *Communications in Mathematical Physics*, 258(2):349–365.

Uffink, J. (2007). Compendium of the Foundations of Classical Statistical Physics. In *Philosophy of Physics*, pages 923–1074. Elsevier.

Uffink, J. (2017). Boltzmann's Work in Statistical Physics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, spring 2017 edition.

Vaidman, L. (2018). Many-Worlds Interpretation of Quantum Mechanics. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall 2018 edition.

Valentini, A. (1991a). Signal-locality, uncertainty, and the subquantum H-theorem. I. *Physics Letters A*, 156(1):5–11.

Valentini, A. (1991b). Signal-locality, uncertainty, and the subquantum H-theorem. II. *Physics Letters A*, 158(1):1–8.

Valentini, A. (1996). Pilot-Wave Theory of Fields, Gravitation and Cosmology. In Cushing, J. T., Fine, A., and Goldstein, S., editors, *Bohmian Mechanics and Quantum Theory: An Appraisal*, Boston Studies in the Philosophy of Science, pages 45–66. Springer Netherlands, Dordrecht.

Valentini, A. (2020). Foundations of Statistical Mechanics and the Status of the Born Rule in de Broglie-Bohm Pilot-Wave Theory. In Allori, V., editor, *Statistical Mechanics and Scientific Explanation: Determinism, Indeterminism and Laws of Nature*. World Scientific.

Valentini, A. and Westman, H. (2005). Dynamical origin of quantum probabilities. *Proceedings of the Royal Society A*, 461:253–272.

Vassallo, A., Deckert, D.-A., and Esfeld, M. (2017). Relationalism about mechanics based on a minimalist ontology of matter. *European Journal for Philosophy of Science*, 7(2):299–318.

Vassallo, A. and Esfeld, M. (2016). Leibnizian relationalism for general relativistic physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 55:101–107.

Villani, C. (2002). A Review of Mathematical Topics in Collisional Kinetic Theory. In Friedlander, S. and Serre, D., editors, *Handbook of Mathematical Fluid Dynamics*, volume 1, pages 71–74. North-Holland, Elsevier Science.

Vlasov, A. A. (1938). On vibration properties of electron gas. *Journal of Experimental and Theoretical Physics*, 8(3):291.

Vlasov, A. A. (1968). The vibrational properties of an electron gas. *Soviet Physics Uspekhi*, 10(6):721.

Volchan, S. B. (2007). Probability as typicality. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 38(4):801–814.

von Neumann, J. (1932). *Mathematische Grundlagen der Quantenmechanik*. Springer, Berlin, Heidelberg.

von Plato, J. (1994). *Creating Modern Probability: Its Mathematics, Physics and Philosophy in Historical Perspective*. Cambridge University Press, Cambridge.

Wagner, G. (2020). Typicality and Minutis Rectis Laws: From Physics to Sociology. *Journal for General Philosophy of Science*.

Wallace, D. (2003). Everett and structure. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 34(1):87–105.

Wallace, D. (2010). Gravity, Entropy, and Cosmology: In Search of Clarity. *The British Journal for the Philosophy of Science*, 61(3):513–540.

Wallace, D. (2012a). *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation*. Oxford University Press, Oxford.

Wallace, D. (2012b). *The Emergent Multiverse. Quantum Theory According to the Everett Interpretation*. Oxford University Press, Oxford.

Wallace, D. and Timpson, C. G. (2010). Quantum Mechanics on Spacetime I: Spacetime State Realism. *The British Journal for the Philosophy of Science*, 61(4):697–727.

Wenmackers, S. (2016). Children of the Cosmos. In Aguirre, A., Foster, B., and Merali, Z., editors, *Trick or Truth? The Mysterious Connection Between Physics and Mathematics*, The Frontiers Collection, pages 5–20. Springer International Publishing, Cham.

Werndl, C. (2013). Justifying typicality measures of Boltzmannian statistical mechanics and dynamical systems. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 44(4):470–479.

Werndl, C. and Frigg, R. (2015a). Reconceptualising equilibrium in Boltzmannian statistical mechanics and characterising its existence. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 49:19–31.

Werndl, C. and Frigg, R. (2015b). Rethinking Boltzmannian Equilibrium. *Philosophy of Science*, 82(5):1224–1235.

Werndl, C. and Frigg, R. (2017). Mind the Gap: Boltzmannian versus Gibbsian Equilibrium. *Philosophy of Science*, 84(5):1289–1302.

Wilhelm, I. (2019). Typical: A Theory of Typicality and Typicality Explanation. *The British Journal for the Philosophy of Science*.

Winsberg, E. (2012). Bumps on the Road to Here (from Eternity). *Entropy*, 14(3):390–406.

Wittgenstein, L. (1922). *Tractatus Logico-Philosophicus*. Kegan Paul, London.

Xia, Z. (1992). The Existence of Noncollision Singularities in Newtonian Systems. *Annals of Mathematics*, 135(3):411–468.