Reductive Aspects of Thermal Physics

Katie Robertson

Pembroke College, University of Cambridge

This dissertation is submitted for the degree of Doctor of Philosophy

October, 2018

Abstract

This thesis examines various reductive case studies in thermal physics. In particular, I argue that according to my account of reduction-as-construction, there are two successful examples of reduction. Thermodynamics reduces to statistical mechanics, and statistical mechanics reduces to the underlying microdynamics — be they quantum or classical. The reduction of a given theory alters that theory's scope, that is: its domain of applicability. The scope of thermodynamics will be central to this thesis — and I will argue for a narrower scope than some authors. This thesis consists of four Chapters, together with an introduction and a conclusion.

In Chapter 1, I discuss how different levels of description relate to one another. I argue that a higher-level of description is reduced to the lower level, if the higher-level quantities and their behaviour can be constructed or captured by the lower-level theory. I claim that 'functionalism' can be helpful in securing reductions. In this Chapter I also argue that the aim of reduction is to vindicate, not eliminate, the higher-level theory.

In Chapter 2, I tackle the reduction of thermodynamics to statistical mechanics. I articulate the functional, or nomological, role of various thermodynamic quantities that are implicitly defined by the zeroth, first and second laws of thermodynamics: temperature, energy and entropy respectively. I then argue that there are quantities in statistical mechanics that realise these roles: though finding them sometimes requires us to focus on quantum, rather than classical, statistical mechanics.

In Chapter 3, I consider the reductive relationship between statistical mechanics and the underlying microdynamics. I demonstrate how the irreversible equations of statistical mechanics can be constructed from the underlying microdynamics using what I label the 'Zwanzig-Zeh-Wallace' framework. Yet this framework uses a procedure called ′coarse-graining′ which has been heavily criticised in the literature; so in this Chapter I offer a justification of coarse-graining. One upshot is that the time-asymmetry in statistical mechanics is weakly emergent.

In Chapter 4, I consider a question about the domain of applicability of thermal physics. Namely: does it apply to self-gravitating systems, such as elliptical galaxies? Much controversy surrounds this question: some argue yes, others argue no. I deflate the dispute by claiming that thermodynamics does not apply, but statistical mechanics does. Thus, my delineation of thermodynamics and statistical mechanics earlier in this thesis not only makes headway with the question of reduction, but also sheds light on this dispute. I argue that this situation — statistical mechanics, but without thermodynamics — can be understood in terms of a central notion in thermal physics: the thermodynamic limit. But as I also discuss: justifying this idealisation has been philosophically controversial.

Preface

This dissertation is the result of my own work and includes nothing which is the outcome of work done in collaboration except as specified in the text. It is not substantially the same as any that I have submitted, or, is being concurrently submitted for a degree or diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. I further state that no substantial part of my dissertation has already been submitted, or, is being concurrently submitted for any such degree, diploma or other qualification at the University of Cambridge or any other University or similar institution except as declared in the Preface and specified in the text. This dissertation does not exceed the word limit of 80,000 words. Chapter 3 is based on a paper forthcoming in the British Journal for the Philosophy of Science. Chapter 4 is based on paper forthcoming in a special issue of Synthese.

Acknowledgments

Firstly, I would like to thank Jeremy Butterfield for his unstinting generosity with his time, wisdom and patience: not least with my total inability to learn how to use a colon rather than a semi-colon. His guidance and friendship has made the experience of writing a thesis —even when trudging through a bog— truly enjoyable.

I owe a large debt of gratitude to many philosophers. Erik Curiel and Owen Maroney have been generous with their time and expertise. I would like to thank Al Wilson for his comments on an earlier draft. For discussion, I would like to thank: Feraz Azhar, Craig Callender, Adam Caulton, Joanna Demaree-Cotton, Natalja Deng, Alex Franklin, Sam Fletcher, Stephan Hartmann, James Ladyman, Casey McCoy, Tushar Menon, Patricia Palacios, Alex Reutlinger, Mathew Simpson, Henry Taylor, David Wallace and Dan Williams.

I am grateful to Pembroke College for the support given to me over the last three years, especially through the Hogwood scholarship. I would also like to thank the Munich Centre for Mathematical Philosophy for welcoming me during a research visit in Michaelmas 2017. Here I benefitted from the kind hospitality of the gemütlich Alma Barner and her philosophical insight over a bowl of chocolate muesli.

I am very fortunate to have two philosophical friends. Carina Prunkl not only read vast swathes of this thesis, but working together, especially in Munich and Obersdorf, tackling the issues of thermal physics, has always been inspirational. I am deeply grateful to Neil Dewar who has been there every step of the way — reading numerous applications, drafts and listening to my half-baked ideas over a cup of coffee. Without his help and friendship, pursuing philosophy of physics would be a lot less fun.

I would also like to thank some non-philosophical friends for their support and encouragement: Charlotte Spires, Grace Edmunds, Louis Devenish, Holly McKenna, Sofia Ropek Hewson, Oliver Burnham and Heather Stallard.

Finally, I would like to thank my family — not only for their enduring love and support, but also for helping me finish this thesis: Dad, for proofreading; Mum, for the sandwiches and clothes; my sister Lizzie (MGIF), for always being there; and my partner Joe, for looking after me during the stressful finishing up. But beyond the thesis, my parents (and the organic carrots) have supported and sustained me over many years and given me every possible opportunity, for which I am forever grateful. I dedicate this thesis to them.

Contents

Introduction 9

1 Wrestling with Reduction 15
1.1 Introduction . . . 15
1.2 List's formal framework . . . 17
1.2.1 Levels of Grain . . . 18
1.2.2 Levels of Description . . . 19
1.2.3 Levels of Dynamics . . . 20
1.2.4 Ontological levels . . . 22
1.3 Supervenience . . . 24
1.3.1 The formal idea . . . 24
1.3.2 The plausibility of supervenience . . . 25
1.4 Multiple realisability . . . 27
1.4.1 Multiple realisability: a problem for reduction? . . . 27
1.4.2 Multiple realisability is widespread — and non-substantive? . . . 29
1.5 The reduction-as-construction account . . . 32
1.5.1 Reduction between what? . . . 33
1.5.2 Reduction-as-construction . . . 34
1.5.3 Dynamical vs. 'non-dynamical' . . . 36
1.5.4 Two kinds of reduction: vertical and horizontal . . . 38
1.6 A tool for reduction: Functionalism . . . 46
1.6.1 Functionalism in the philosophy of physics . . . 46
1.6.2 Functionalism in the philosophy of mind . . . 47
1.6.3 Functionalism fit for physics . . . 49
1.6.4 Conclusion . . . 52
1.7 Metaphysics and reduction . . . 52
1.7.1 Emergent special sciences . . . 54
1.8 Reduction in practice vs in principle . . . 55
1.9 Conclusion . . . 57

2 The reduction of thermodynamics to statistical mechanics 75
2.1 Introduction . . . 75
2.1.1 Silencing scepticism . . . 76
2.1.2 Autonomy . . . 77
2.1.3 Prospectus . . . 79
I. Thermodynamics in general . . . 80
2.2 Equilibrium state-space . . . 80
2.3 Dynamics . . . 81
2.4 Thermodynamics as control theory: Interventions . . . 85
2.5 Interventions and anthropocentrism . . . 86

2.6 Statistical mechanics construed . . . 87
    2.6.1 Gibbs vs. Boltzmann . . . 88
    2.6.2 Introducing probability . . . 89
    2.6.3 The Workhorse . . . 91
2.7 The differing concerns of SM and TD . . . 93

II. The Zeroth Law . . . 93

2.8 Pure thermodynamics: the functional role of temperature . . . 94
2.9 The statistical mechanical realiser . . . 95
2.10 Batterman's reconstruction of Gibbs' objection . . . 97
    2.10.1 The connection to the Canonical Ensemble . . . 99
2.11 Temperature is not mean kinetic energy . . . 100
    2.11.1 Derivation of the ideal gas law . . . 100
    2.11.2 "$T \neq \langle K \rangle$" in general . . . 102
2.12 Conclusion . . . 104

III. The First Law . . . 104

2.13 Phenomenological Statement of the First Law . . . 105
2.14 A historic achievement . . . 106
2.15 Done and dusted? . . . 107
2.16 Molecules in motion . . . 109
    2.16.1 Macro/micro . . . 109
    2.16.2 Anthropocentrism? . . . 110
    2.16.3 More than molecules in motion . . . 111
2.17 The image in SM: quantum heat and work . . . 112
2.18 The First Law: Conclusion . . . 115

IV. The Second Law . . . 116

2.19 Introduction . . . 116
2.20 Concepts of Irreversibility . . . 117
2.21 The source of all asymmetry? . . . 118
2.22 The Second Law Introduced . . . 119
    2.22.1 Carnot Cycle . . . 120
    2.22.2 Defining Entropy . . . 121
2.23 Three key features of the TDSL . . . 123
    2.23.1 The importance of the environment . . . 123
    2.23.2 Quasi vs. non-quasi static processes . . . 124
    2.23.3 The Second Law vs. the Minus First Law . . . 124
2.24 Brownian motion, the Atomic Hypothesis, and the Demonic consequences... . . . 125
    2.24.1 Brownian motion . . . 126
    2.24.2 Loschmidt's demon . . . 127
    2.24.3 Maxwell's demon . . . 128
2.25 Four different views about the scope of the TDSL . . . 130
2.26 Connections between options 1.-3. . . . 131
2.27 The Demons and the Maxwellian means-relative view . . . 134
    2.27.1 The difference between Maxwell's demon and Loschmidt's demon: resisting the Maxwellian view . . . 135
2.28 Statistical Mechanics and Maxwell's demon . . . 136
    2.28.1 Thermodynamics and possibility . . . 136

2.28.2 The explanandum 137
2.28.3 Details of the demon: controlled operations 137
2.29 Landauer's principle 139
2.29.1 Why the demon needs a memory 140
2.29.2 Why the demon needs to reset its memory 141
2.29.3 How reset leads to entropy increase 141
2.29.4 The Status of Landauer's principle 143
2.30 Conclusion about the scope of the TDSL: the verdict on options 1-4. . . . 144
2.31 The Image of the TDSL in SM 145
2.31.1 The Strife and the Holy Grail 146
2.31.2 Non-decreasing SM entropy is neither necessary nor sufficient . . 147
2.31.3 Quasi-static changes in SM 147
2.32 Conclusion 151

V. The Third Law 152

2.33 Conclusion: TD reduces to SM 154
2.34 Appendix: Gibbs and Boltzmann entropy 155

3 Justifying coarse-graining in statistical mechanics 158
3.1 Introduction 158
3.1.1 Prospectus 159
3.2 The ZZW framework 159
3.3 Why does this method work? 165
3.3.1 The special conditions account 166
3.3.2 When is a density forwards-compatible? 167
3.4 Anthropocentrism and illusion: two objections 170
3.4.1 The two objections in more detail 171
3.4.2 Against the justification by measurement imprecision (MI) 173
3.5 An alternative justification 174
3.5.1 Abstraction and autonomy 175
3.5.2 An illustration: the Game of Life 178
3.6 Reply to (*Illusory*) 180
3.7 Reply to (*Anthropocentric*) 181
3.8 The wider landscape: concluding remarks 182
3.8.1 Inter-theoretic relations 182
3.8.2 The nature of irreversibility 183

4 Stars and steam engines 187
4.1 Introduction 187
4.2 Newtonian Gravity weighs in 189
4.2.1 How the situation changes with gravity 190
4.2.2 Successes 191
4.2.3 Unusual features 193
4.3 My Strategy for Reconciliation 196
4.4 Thermodynamics "Construed" 200
4.4.1 Thermodynamics in general 200
4.4.2 Thermodynamics does not apply to SGS 202
4.5 Does statistical mechanics apply to SGS? 204
4.5.1 An approach to equilibrium? 204

4.5.2 Surprise? 206
4.6 The Bottom-Up Explanation 206
4.6.1 The connection to reduction 211
4.7 Conclusion 212

Conclusion **213**

Introduction

The reduction of thermodynamics to statistical mechanics is a central, and classic topic in philosophy of science. But one might wonder: is it a tired topic? One might hold that for many years, surely since e.g. Nagel (1961), it has been clear that thermodynamics successfully reduces to its successor, statistical mechanics. Yet historically, it was far from obvious that thermodynamics would reduce to a lower-level theory or indeed that it would not itself be foundational. Thus, Kragh (2001) summarises the debate in the 1890s by saying: "thermodynamics.... was sometimes argued to not only be different from mechanics in principle, but also to have priority over mechanics as a more satisfactory foundation on which all of physics could be built" (Kragh, 2001, p. 7). Moreover, in recent years, there has been a resurgence of interest in the relationship between thermodynamics and statistical mechanics. This literature has centred on phase transitions, where an infinite limit must be invoked in order to recover the thermodynamic description in statistical mechanics (cf. inter alia (Batterman, 2001), Menon and Callender (2011), Butterfield and Bouatta (2012)). Consequently, there has been much scepticism in the philosophy of physics over whether thermodynamics reduces to statistical mechanics.

Not only is the relationship between thermodynamics and statistical mechanics of philosophical interest, it is also of real interest in theoretical physics. It is important for the current quest for a theory of quantum gravity. The thermodynamics of black holes guides the search for a theory of quantum gravity in a specific way: the hope is that black hole thermodynamics will be to a theory of quantum gravity, as thermodynamics is to statistical mechanics (Wall, 2017). As such, the thermodynamics-statistical mechanics relationship is meant to be analogous to the relationship between black hole thermodynamics and quantum gravity. But such a claim obviously prompts philosophers of physics to ask: what exactly is the relationship between thermodynamics and statistical mechanics?

This question is at the heart of this thesis. But fascinating though the case studies of phase transitions and black holes are, there is much philosophical work to be done in central but less popular areas. The term 'thermal physics' conjures up a cluster of theories: the kinetic theory of gases, phenomenological thermodynamics, classical and quantum statistical mechanics as well as condensed matter physics. Here I will be concerned with three different levels of description, described by:

1. Thermodynamics.
2. Statistical mechanics.
3. The underlying microdynamics: classical mechanics, and quantum mechanics.

Overall, I will claim: there are two cases of successful reduction in thermal physics. (1) thermodynamics reduces to statistical mechanics, and (2) statistical mechanics reduces to the underlying microdynamics.

Proclamations of successful reduction are, of course, relative to an account of reduction. So throughout this thesis I develop my views on reduction. Of course, the word 'reduction' looms large in other areas of philosophy. Metaethics addresses the question: do normative properties reduce to non-normative properties? And in philosophy of mind perhaps the most central question of all is: does the mind reduce to the brain? That is, does the mental reduce to the physical? This latter question of course intersects with the philosophy of science especially in the debate about whether the special sciences, such as psychology, reduce to physics.

This literature about the special sciences treats 'physics' as a homogenous enterprise. But there are many different levels of description within physics — even beyond the three levels that this thesis focuses on. Thus, the types of issues that arise in the debate about the special sciences — multiple realisability, emergence and autonomy — arise also *within* physics. Indeed, there is good reason to think that thermodynamics qualifies as a special science (cf. Hemmo and Shenker (2015)). After all, one crucial feature of the special sciences is that they are non-fundamental.

Yet thermodynamics is an unusual physical theory, in various ways. As a recent review jokingly put it: "If physical theories were people, thermodynamics would be the village witch.... The other theories find her somewhat odd, somehow different in nature from the rest, yet everyone comes to her for advice, and no one dares to contradict her" (Goold et al., 2016, p. 1).

One way in which thermodynamics differs from other physical theories is that there are no spontaneous dynamics under which the system evolves — independently of our description of, or interest in, the system. Instead, thermodynamics concerns the interventions that we can or cannot perform on a system. For this reason, Planck worried that thermodynamics is anthropocentric (cf. Uffink (2001)). Furthermore, some claim that the time-asymmetry in thermal physics is also anthropocentric — in a way that prevents one from being a scientific realist about these theories. Throughout this thesis, I will defuse this worry, by arguing that theories in thermal physics are not *inherently* anthropocentric, in any way that is different from the rest of our scientific theories.

Another general theme of my approach is what I call a 'pragmatic naturalism'. By this I mean: the practice of science must be given priority over philosophers' quibbles.

That this is my background philosophical position is manifest in several places. For example, I claim that the aim of reduction is to vindicate, rather than eliminate, the higher-level theory and its entities. The only reason to eliminate a theory is that it is not longer useful. This is why even if we had a completed physics, this would have a limited impact on the special sciences.

This pragmatic naturalism is also visible in my focus, in my account of reduction, on the practice of science. Thus, I focus on the mathematical and conceptual relationships that in fact obtain between theories, rather than focussing on questions of how they could in principle be related. Approximations, idealisations and abstractions are a crucial part of the practice of physics— and accordingly, their importance has often been emphasised by anti-reductionists, such as Batterman (1995). I hope to incorporate these important insights into my account of reduction. Anti-reductionists also emphasise the importance of the higher-level theory — and this supports the anti-eliminative view. Furthermore, the higher-level theory is often 'autonomous' from the lower-level details — at least, to a certain extent. This robustness, or imperviousness to the lower level, is especially prominent in the case of thermodynamics: the theory makes a lot of progress in blissful ignorance of the nature of matter.

I hope that my account of reduction throughout this thesis can capture some of the insights of both the reductionist and anti-reductionist camps. Consequently one might worry: am I giving an account of reduction worth the name? I contend that the answer is yes. After all, an anti-reductionist position is defined in opposition to a particular conception of reduction. If my account is one that self-describing anti-reductionists do not wish to oppose, then I take this to be progress.

In Chapter 1, I outline my account of reduction and the relationships between different levels of description. I outline, and endorse, List's formal framework of levels. After discussing the plausibility of supervenience and the relationship to reduction, I set out my view of reduction-as-construction. According to this view: the higher-level theory T_t (t for 'top') is reduced to the lower-level theory T_b (b for 'bottom') if the equations and quantities of T_t can be constructed — using whichever approximations, idealisations and abstractions the physicist requires — from T_b.

I then explore why advocating 'functionalism' has in recent years become a more popular position in philosophy of physics. I claim that this popularity is because functionalism can calm some worries about differences between the different levels of description. But I also claim that an account of functionalism fit for physics differs from some accounts of it in the philosophy of mind. Functionalism will be especially central in Chapter 2, where I consider the case study of thermodynamics and statistical mechanics.

In Chapter 2, in order to tackle the relationship between thermodynamics and statistical mechanics, I have to first construe both theories. This is in itself a substantive project, because in the practice of physics the concepts are often blurred together. Stipulating the content of the two theories is a substantive project partly because thermal physics is a strange beast, untamed by the conceptual rigour that other physical theories possess. In particular, statistical mechanics is often considered to be a collection of frameworks and distinct schools of thought, with no uncontroversial formalism (e.g. Uffink (2006a, p. 4)). As a result, there are many conceptual problems and approaches that I have to leave to one side in this thesis.

The relationship between thermodynamics and statistical mechanics is a vast topic, and consequently I cannot tackle all of the central issues in a sustained and collected manner. Two conspicuous omissions will be the nature of both probability and entropy. Much controversy surrounds the nature of entropy: Jaynes and Wigner called entropy 'anthropomorphic' (Jaynes, 1965, p. 6); Lloyd (2006), Bekenstein (1973) and others claim it is an information-theoretic concept. Some evangelically advocate and others vehemently oppose the connection between entropy and ignorance, or our state of knowledge cf. Denbigh and Denbigh (1985). Whilst I believe that my rejection of anthropocentrism in thermodynamics and statistical mechanics throughout this thesis will have ramifications for the nature of entropy, I do not explore them here.

In part, this is because considering the nature of entropy involves a more sustained engagement with the topic of probability in statistical mechanics than I am able to give here. But, as I discuss when 'construing' statistical mechanics, I believe statistical mechanics probabilities are wholly objective. Furthermore I believe — although I cannot argue for this here — that the problem of understanding probability in statistical mechanics cannot be disentangled from the problem of understanding probability in QM (cf. Wallace (2016)). Indeed, understanding the nature of probabilities in statistical mechanics is one of the two main foundational problems in statistical mechanics. The second of which I consider in Chapter 3: the nature of time-asymmetry.

Having construed statistical mechanics and thermodynamics, I focus on reduction. I claim that we can consider the functional, or nomological, roles of the quantities implicitly defined by the laws of thermodynamics and then search for the quantities in SM that realise these roles. Whilst traditionally functionalism has involved the entire theory, my strategy will be to take the laws of thermodynamics one by one; I proceed from the Zeroth Law to the Third law articulating the functional roles of the quantities, temperature, energy and entropy respectively; and then finding the realisers in SM.

In the course of this project, I will make several substantive (i.e. controversial!) claims, such as:

- I will claim that — in general — temperature is not mean kinetic energy.
- I will claim heat and work do not simply correspond to 'disordered' and 'ordered' molecular motion.

- I will claim that the Second Law is not quite the titan it is sometimes claimed to be. For instance, Atkins (2007) goes as far as to say that the Second Law is the reason 'anything at all' happens. By distinguishing the different types of irreversibility, and by distinguishing the Minus First Law from the Second Law, I will show that the scope of the Second Law is constrained to cases where quasi-static processes are possible.

- I will find the realiser of the Second law within *quantum* statistical mechanics. As such, I break with the tradition in the philosophy of thermal physics by not limiting myself to classical statistical mechanics: in certain cases, quantum considerations are important (as emphasised by Wallace (2015c,a) and Albert (2000)).

By considering the relationship between thermodynamics and statistical mechanics in this functionalist manner, the absence of the thermodynamic limit in my discussion may seem conspicuous. The thermodynamic limit is, roughly speaking, the idealisation that the number of constituents of the system is infinite. Justifying this idealisation is philosophically controversial: no actual systems are infinite, and yet certain features only obtain in this limit — and, worryingly, our descriptions of these features seem resistant to 'de-idealisation'. Usually, the idea is that the exact features of the TD description are recovered from the SM description in this limit — as will be discussed in Chapter 4. But we can make headway with the project of reduction without considering the thermodynamic limit. Indeed, seeing this done in Chapter 2 and 3 should be a welcome advance, given the worries about no actual systems being infinite. Thus, I delay my discussion of the thermodynamic limit until Chapter 4.

In Chapter 3, I consider the relationship between statistical mechanics and its underlying microdynamics. This relationship has been considered paradoxical or puzzling. The equations of statistical mechanics are time-asymmetric but the microdynamics which underpin these equations are time-symmetric. Where does the asymmetry come from? How can the two levels be reconciled?

I argue that progress can be made by considering what I term 'the Zwanzig-Zeh-Wallace' framework. This framework constructs the time-asymmetric equations of statistical mechanics from the underlying dynamics, but in doing so requires an initial condition assumption. This is where the motivation for considering a 'Past Hypothesis' comes from: but for the most part I leave such cosmological considerations to one side (cf. Wallace (2011)). This framework also uses a procedure that has been heavily criticised: coarse-graining. Thus, the main project of this Chapter is to give a justification of coarse-graining.

According to my account of reduction developed in Chapter 1, the two levels at issue are reconciled with one another in this framework: statistical mechanics is reduced to the underlying microdynamics, be they classical or quantum. Besides, in Chapter 1 I will argue that my account of reduction is compatible with emergence. And this compatibility is illustrated by the case study of Chapter 3: the time-asymmetry, i.e. irreversibility, of statistical mechanics is weakly emergent.

In Chapter 4, I consider the scope of thermal physics. In particular, does it apply to self-gravitating systems such as elliptical galaxies or globular clusters? These systems contain roughly between 10^5 and 10^{11} stars interacting by Newtonian gravity. Callender (2011) raises this topic by asking: does the distribution of stars in the night sky have a thermal explanation? There is a live dispute over this question: some physicists argue yes, others argue no.

Thermodynamics makes few assumptions about the constitution of the system. Consequently, many, such as Planck and Clausius, have argued that TD has universal scope (Uffink, 2001). Thus, Einstein is quoted as saying: " [Thermodynamics] is the only physical theory of universal content, which I am convinced, that within the framework of applicability of its basic concepts will never be overthrown" (Klein, 1967, p. 509).

But I resist this claim of universal scope. I argue that the dispute over the applicability of thermal physics to self-gravitating systems can be deflated if we properly distinguish thermodynamics and statistical mechanics. Here, my delineation of thermodynamics and statistical mechanics echoed from Chapter 1, bears fruit.

I claim that thermodynamics does not apply to self-gravitating systems — as these systems do not display the right equilibrating behaviour and so there is no equilibrium state space appropriate for these systems. Here, the functionalism about thermodynamics argued for earlier comes into play: whilst thermodynamics itself does not legislate whether it applies to stars or steam engines, the system in question has to display certain behaviours so that there are quantities to fill the thermodynamic roles: and self-gravitating systems do not play the right role.

But I argue that whilst thermodynamics is inapplicable, there is — to a certain extent — a statistical mechanical description of self-gravitating systems. This case study demonstrates that even though thermodynamics reduces to statistical mechanics, the scope of applicability of the two theories can still come apart. Besides, I claim that the (non-existence) of the thermodynamic limit can explain this difference of scope: statistical mechanics without thermodynamics.

1 Wrestling with Reduction

1.1 Introduction

Nature admits of different descriptions. Eddington (1928) famously described a table in radically different ways: it can be described as wooden and solid, or alternatively as consisting of concentrated masses whose diameter is far smaller than the chasm between them. Leaving aside the differences between the descriptions of the manifest and scientific images of the world (Sellars, 1963), even within science there are a multitude of different descriptions. And like Eddington's table, which description is appropriate depends on the context.

Understanding how different descriptions given by different scientific theories fit together, if indeed they do, is the project of inter-theoretic reduction. The seminal work of Nagel (1935, 1961) (and in his footsteps, Schaffner (1967)) has defined much of the subsequent literature on inter-theoretic reduction in science. This account is in keeping with the philosophy of its time; Nagel's contemporary Hempel (1948) advocated the covering law account of explanation. An explanandum was successfully explained if one could provide an argument with an explanans that invokes the appropriate laws of nature as a premise, and from which one can logically deduce the explanandum. Nagel's account has a kindred explanatory spirit: to reduce the higher-level theory (henceforth: T_t) to the lower-level theory (T_b) one must be able to derive the laws of T_t from the laws of T_b, in certain cases with the help of appropriate 'bridge laws'.[1] A bridge law defines the terms, or vocabulary, of the higher-level theory in terms of the terms of T_b, where the definiens is on the right hand side of the universally quantified biconditional.

Since Nagel, there has been much debate about the correct account of reduction. In this Chapter I cannot do justice to that rich philosophical heritage, nor do I aim to give an account of reduction appropriate to all possible cases. But in what follows I will argue for a particular account of reduction, which will be the standard to which my reductive claims throughout this thesis will be held.

The ubiquity of stable, macroscopic patterns means that it is uncontroversial in science that sometimes the most productive way to describe a given phenomenon is to

[1] Whether bridge laws are required or not defines whether the reduction is inhomogeneous or homogeneous respectively. But the details of Nagel's account will not be required in what follows.

discuss it at the 'higher-level of description' rather than in terms of its more fundamental constituents. Little progress would be made in ecology if the only permissible language was that of physics. Examples abound: understanding why platypuses have venomous spurs and no teeth, is a challenging task from the perspective of evolutionary biology, but asking for a description in terms of cellular biology—let alone quantum mechanics—is surely an unreasonable request. Different levels are appropriate for different tasks. And so the talk of 'levels of description' fits naturally with scientific practice.

But understanding the relationships between these levels is a controversial —yet central— philosophical project. Why is asking for a description of a platypus in terms of the Schrödinger equation an unreasonable request? Is the higher level an imperfect way of grasping the lower level? A mere crutch or heuristic we rely on, owing to our cognitive limitations? Or does reality itself admit of different levels, perhaps in the manner described by the British Emergentists (Broad, 1925; Alexander, 1920)? Or is 'levels talk' merely metaphorical?

In Section 1.2, I discuss and endorse a formal framework given by List (2017), which makes levels talk precise. One assumption of this framework is that different levels are related by supervenience mappings; I explicate supervenience and its philosophical significance in Section 1.3.

One feature of List's framework is that the higher-level worlds are multiply realised by the lower-level worlds. Traditionally, multiple realisability has been taken to be problematic for certain accounts of reduction. In Section 1.4 I rebut these objections in the literature (following Sober (1999)). I then suggest that multiple realisability is widespread. Shapiro (2000) resists this claim by distinguishing between substantive and trivial cases of multiple realisability, but I argue that this move is unsuccessful. Admittedly, some putative cases of multiple realisability may be surprising — but part of the reductive project is discovering which differences do not matter — and so I will argue that reduction renders seemingly surprising examples of multiple realisability unsurprising.

In Section 1.5, I outline my account of reduction-as-construction. I will claim that the higher-level theory (T_t), or a model of this theory is reduced to the lower-level theory (T_b), or a model of that theory, if the equations or quantities of T_t can be *constructed* from the equations or quantities of T_b. Because the crucial aim of reduction is that the *behaviour* of the phenomena described by T_t is captured by T_b, I claim, in Section 1.6, that in some cases functionalism can be helpful in securing reduction. The quantities of T_b need to fill the role of the quantities described by T_t — but not all the features of T_t will be part of this role.

Whilst I will claim that functionalism is sometimes helpful, I will reject Kim (1998, 1999)'s account which focuses on causal roles — not least because it is opaque how to understand causation in physics. I also reject the eliminative aim of his account.

Instead I argue in Section 1.7 that if T_t is reduced to T_b, this is a *vindication* of T_t. We should only eliminate the properties described by T_t if they are no longer useful — and I will give various reasons why a reduced theory is often still useful. In this way, the considerations about reduction are sometimes orthogonal to our metaphysical commitments. Further, I argue that reduced higher-level descriptions might earn the name 'emergent'. Finally, in Section 1.8, I discuss the difference between reduction in principle and reduction in practice, and why I am concerned with reduction in practice; not least because the rest of this thesis is concerned with case studies.

1.2 List's formal framework

Whilst 'levels-talk' is popular, is it just a mere metaphor? (cf. Kim (2002) Owens (1989)). In this Section, I outline how levels-talk is made precise by a framework suggested by List (2017) (and ancestors, or close cousins of this framework are proposed by many authors, inter alia Lewis (1988) and Butterfield (2012)). List outlines a formal framework for considering 'systems of levels'. A system of levels is a pair $< \mathcal{L}, \mathcal{S} >$:

- $\mathcal{L}$ is a class of objects called 'levels'.
- $\mathcal{S}$ is a class of mappings between the different levels, called 'supervenience mappings'. Such a mapping σ goes from the source ('lower') level L to the target ('higher') level L', $\sigma : L \to L'$. For any given pair of levels, there is at most one map in S.

Another crucial feature of σ is that it is a *surjective function*. There is a set of possible worlds Ω at each level. This means that each world $w \in \Omega$ at the source level L gets mapped to *at most* one world $w' \in \Omega'$ at the target level, L': thus, σ is a *function*.[2] *Surjectivity*: every world w' at the target level L' is in the range of the σ function. This means that no higher-level possible world lacks a lower-level realiser: we can write this as 'w is the lower-level realiser of w''.

In addition to discussing worlds at one level supervening on worlds at another level, we can discuss facts at one level supervening on facts at other. The propositional content (i.e. intension) of a sentence ϕ is the set of worlds where that sentence is true. Thus, the propositional content of a sentence asserting a certain fact is the set of possible worlds where that fact obtains. In List's words: "Let $E' \subseteq \Omega'$ represent some higher-level fact, namely the fact that the higher-level world falls inside the set E'" (List, 2017, p. 11). $\sigma^{-1}(E')$ is the inverse image under the supervenience mapping:

[2] Strictly speaking, σ must be a partial function: not every world at the source domain need be mapped to *any* world at the target level: for example, not all physically possible worlds lead to biologically possible worlds. My thanks to Sam Fletcher for this point.

$\sigma^{-1}(E') = \{w \in \Omega : \sigma(w) \in E'\}$. This set of lower-level worlds is the supervenience basis of E'. $\sigma^{-1}(E')$ is called the lower-level fact corresponding to E'.[3]

List uses the terminology of 'levels' but points out that 'scale', 'domain' and 'subject matter' are also apt: the latter two especially so, because it makes it clear that the levels need not be linearly ordered, but only partially. Thus, 'subject matters' or 'domains' makes it clear that geology and biology might both supervene on the physical, while neither biology nor geology supervenes on the other.

A system of levels has certain formal features so that it forms a mathematical object: a category. The resources of category theory, in particular the different types of maps known as functors, can be brought to bear on the relationships between different systems of levels. List considers four different types of systems: levels of grain, ontological levels, levels of description, and levels of dynamics, which I take up in turn in what follows.

Note that List is operating at a greater level of generality in considering the relationships between *systems* of levels. When considering inter-theoretic reduction, we are concerned with the relationships between certain different levels *within* one system of levels. Indeed, the examples most appropriate for discussing inter-theoretic relations between scientific theories is List's 'levels of description' and 'levels of dynamics', which I explicate in Section 1.2.2 and Section 1.2.3, after, in Section 1.2.1, outlining the levels of grain. Then in Section 1.2.4, I consider how these systems of level relate to each other and ontological levels.

1.2.1 Levels of Grain

There is one set of possible worlds Ω. This set can be partitioned in different ways, by different equivalence relations. (In effect, this is the same apparatus as Lewis (1988)'s subject matters[4]). $\Omega_\sim$ denotes the set of equivalence classes induced by the $\sim$ relation. If $\sim$ is just the identity relation, then this is a finest-graining possible. If $\sim$ is the total relation so that every world in Ω is in the same equivalence class, then this is the coarsest-graining possible. Of course, the interesting cases are somewhere in between these two extremes.

We can generate a system of levels as follows. For any two partitions $\Omega_\sim$ and $\Omega_\approx$, we say that $\Omega_\sim$ is at least as fine-grained as $\Omega_\approx$ if each equivalence class in $\Omega_\approx$ is a

[3] List (2017) advocates a world-based, rather than entity-based, understanding of levels. In an entity-based understanding the 'higher than' relation is not modal (as supervenience is) but mereological. But the part/whole relation doesn't always line up nicely with the higher/lower than relation. Instead, once we have the best theory of a particular level, we can then commit to those level-specific entities. See (List, 2017, p. 11), Norton (2014), Block (2003) for more details on this debate.

[4] Note, however, that for Lewis (1986b) the worlds are real possibilities, but for List, the worlds are sets of sentences. When considering levels of description which each level has a language **L**, the worlds are defined as maximally consistent sets of sentences of the language. However, the details and status of possible worlds is not needed for what follows.

union of equivalence classes in $\Omega_{\sim}$. The relation "is at least as fine-grained as" can then be used to partially order partitions. And in these cases we can define a function $\sigma : \Omega_{\sim} \rightarrow \Omega_{\approx}$, which assigns to each equivalence class in $\Omega_{\sim}$ the equivalence class in which it is included in $\Omega_{\approx}$.

List gives an example from decision theory: an agent's possible levels of awareness can be modelled as levels of grain. An agent has a greater level of awareness if they can draw finer distinctions between different possibilities. Such levels of grain are often understood epistemically, as different ways of representing the world. Here is an example from physics: levels of precision. The outcomes of repeated experiments can be put into different equivalence classes, and there will be different partitions depending the level of precision required. Jane and Michael might have found the same value for the gravitational constant, g — depending on how precise one is. Two outcomes $g = 9.81$ and $g = 10$ will be put in the same equivalence class if the level of precision is 'to the nearest natural number': but not according to a finer-grained and so more precise partition.

1.2.2 Levels of Description

In order to define a system of levels of description, List defines a formal language **L**, which has certain features (a negation operator and a notion of consistency): an example of such a language is standard propositional logic. Defining a language **L** induces an *ontology*, which is a minimally rich set of worlds Ω_L such that (i) each world in Ω_L 'settles' everything that can be expressed in **L** and (ii) nothing else is settled that is not entailed by what is expressible in **L**. Different languages **L** and $\mathbf{L}'$ generate two sets of worlds Ω_L and $\Omega_{L'}$ respectively. We define a system of levels of description as the pair $< \mathcal{L}, \mathcal{S} >$, where each level L in $\mathcal{L}$ is a pair $< \mathbf{L}, \Omega_L >$ and S is a class of surjective functions of the form $\sigma : \Omega_L \rightarrow \Omega_{L'}$.

The connection to inter-theoretic reduction is that a scientific theory can be considered to be a set of sentences in a particular language.[5] Different sciences use different predicates: physics uses the predicate 'is an electron' whereas as evolutionary biology uses the predicate 'is a platypus'. Of course, the languages invoked by List are formal languages; the reality of science in practice is much more messy. Whilst this idealised formal approach is helpful for setting the scene, I will ultimately defend a pragmatic account of reduction.

[5] This is the syntactic view of theories, but see Halvorson (2013), Lutz (2017) and Hudetz (2017), for a discussion of the view that the syntactic and semantic views of theories are equivalent.

1.2.3 Levels of Dynamics

Another system of levels that List (2017) and Butterfield (2012) describe are levels of dynamics. Coin flipping can be modelled at a detailed microscopic level: think of calculating all the forces and initial position and the exact angle. This level describes the coin's trajectory. Alternatively, the coin flip can be modelled on a coarser probabilistic level, where there are only two outcomes 'heads' and 'tails' (cf. (List, 2017, p. 15-16)). Likewise, the weather, the economy and the climate can either be studied at a 'micro-dynamical' or a 'macro-dynamical' level.

To define a dynamical system, let us first define the lower-level state space $\mathbb{S}$. An element $s \in \mathbb{S}$ represents a possible state of the system, which is an assignment of a value of each quantity. Following Butterfield (2012, p. 15), the set of quantities $\mathbb{Q}$ varies from theory to theory. But as an intuitive example, one can imagine that each quantity $Q \in \mathbb{Q}$ is a real-valued function on $\mathbb{S}$. We can think of the differences in values of certain quantities as distinguishing the different microstates. That is, for two distinct elements $s_1 \neq s_2 \in \mathbb{S}$, there is a quantity $Q \in \mathbb{Q}$ such that $Q(s_1) \neq Q(s_2)$.

Next we can define the lower-level dynamics as a map $D : \mathbb{S} \rightarrow \mathbb{S}$. The dynamics could be one-to-one, in which case we called it deterministic. If the dynamics are one-to-many, they are 'past-to-future' indeterministic. Finally, if they are many-to-one, they are 'future-to-past' deterministic. A history h is a map from the linearly ordered instants of time $t \in T$ into $\mathbb{S}$, $h : T \rightarrow \mathbb{S}$. For each time $t \in T$, a state of the system $s \in \mathbb{S}$ is assigned. The set Ω of possible histories allowed by the dynamics is a set of nomologically possible histories, and this set Ω plays the role (at this lower level) of the set of possible worlds discussed above.

We can now define the higher-level state space. In keeping with the above notation, I write upper dashes to denote the higher-level state space $\mathbb{S}'$ and higher-level quantities $\mathbb{Q}'$. We assume that the higher level supervenes on the lower level. For example, the higher-level states could be aggregates, or equivalence classes, of the lower level. For example, in the coin-tossing case the higher-level states are 'heads' or 'tails'. The higher level is the macroscopic level — and a macrostate is compatible with many different lower-level microstates. One example: the aggregate state of a glass of water is compatible with many different configurations of the water molecules within the glass.

There is a partition P on $\mathbb{S}$ that splits the totality of microstates into exhaustive and mutually exclusive subsets $C_i \subset \mathbb{S}$. The cells of the partition represent the macrostates $s' \in \mathbb{S}'$. The supervenience map σ takes each microstate s to a macrostate s', as before: we require that $\sigma : \mathbb{S} \rightarrow \mathbb{S}'$ is a surjective function. The function will induce a mapping from the set of micro histories Ω to the set of macro histories Ω'. (Here, a macro-history h' is of course defined as an assignment of a macrostate s' to each time, $t \in T$.)

We can consider the induced macrodynamics D' as follows. We already have the microdynamics $D : \mathbb{S} \to \mathbb{S}$, we now want to consider how the macrostates s' evolve over time: i.e. what is $D' : \mathbb{S}' \to \mathbb{S}'$. We know that each macrostate $s' \in \mathbb{S}'$ is a set of microstates $C_i \subset \mathbb{S}$. Each state s in the C_i evolves under the microdynamics. What is the temporal evolution of a cell C_i, $D(C_i)$? Generally, the image of $D(C_i)$ is not another single cell, i.e. macrostate: $s_1 \neq s_2 \in C_i$ might get sent to microstates s_3 and s_4 which are elements of distinct cells C_j and C_k. Thus, the induced macrodynamics can be indeterministic, even if the microdynamics is deterministic. This is illustrated in Figure 1.1.

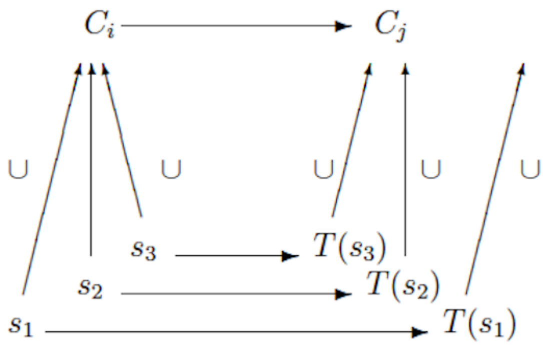

Figure 1.1: The failure of meshing, from Butterfield (2012). Here $\bigcup$ represents σ, and T represents D.

But if the microdynamics takes every state in the partition cell C_i to another cell C_j, i.e. the image $d(C_i)$ is another element of the partition, then the microdynamics induces deterministic macrodynamics: for which an appropriate jargon proposed by Butterfield (2012) is 'meshing dynamics'. In other words, the time evolution D and coarse-graining implemented by σ commute. See Butterfield (2012) for a range of interesting examples of successful and unsuccessful meshing.

Thus, dynamics at different levels can be interestingly different from one another. List and Pivato (2015) give an example of how the microdynamics can be deterministic (i.e. one-to-one), but the induced macrodynamics can be indeterministic, i.e. a failure of meshing as in Figure 1.1. This is vividly illustrated in Figure 1.2.

The levels of dynamics will be especially useful for considering cases of reduction in physics. One important difference to note about levels of dynamics in physics, compared to more speculative cases of reduction between special science and physics, is as follows. In the case of levels of dynamics, it is clear that the same system is being described from different perspectives. ('System' is, of course, just scientific jargon for the philosopher's 'object'). But it is not obvious that the same 'system' will be under discussion when a higher-level description E' is pulled backed under σ^{-1}. Indeed, the

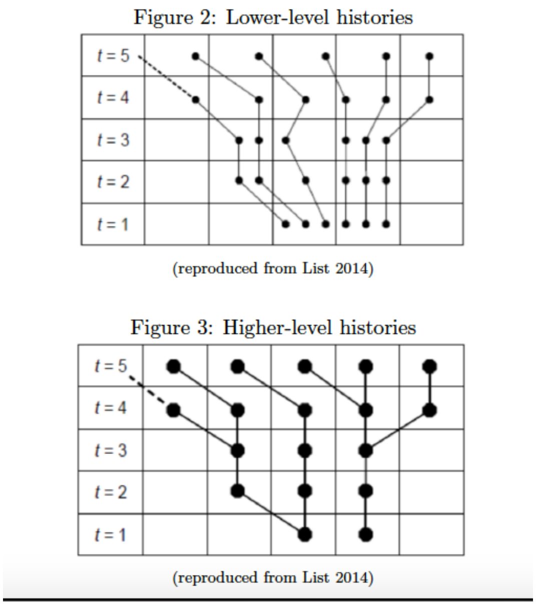
Figure 2: Lower-level histories

(reproduced from List 2014)

Figure 3: Higher-level histories

(reproduced from List 2014)

Figure 1.2: Emergent chance: deterministic lower-level histories are compatible with higher-level indeterminism, List and Pivato (2015).

lower-level description E may not be recognisably about the same object.

1.2.4 Ontological levels

Finally, we have List's systems of ontological levels. Not only do our representations of reality stratify into different levels, but reality itself is stratified into levels. Instead of just having one set of possible worlds (over which we consider different partitions), there are now different worlds at different levels. Level-specific worlds can be viewed as specifications of level-specific properties. A possible world at a particular level is a full specification of the way the world might be *at that level*. For example, the possible worlds at the social level is a specification of the totality of social facts (List, 2017, p. 9), whereas worlds at the chemical level require a specification of the totality of chemical facts. "From a lower-level perspective, a higher-level world thus looks like a partial or incomplete specification of the world, which leaves certain facts (namely lower-level ones) indeterminate" (List, 2017, p. 9).

One feature relevant for future Sections is that "the relationship between higher-level and lower-level worlds is one of supervenience with multiple realisability" (List, 2017, p. 9). The lower-level facts fix the higher-level facts: the physical facts 'fix' the chemical facts, for instance. Each physical world $\omega \in \Omega$ gets mapped by a σ function to a chemical world $\omega' \in \Omega'$: ω is physical realiser of ω'. The supervenience requirement is that σ is surjective — there is no higher-level world which lacks a lower-level realiser. But one particular chemical world could be compatible with many physical worlds:

that chemical world is multiply realised. In other words, the supervenience map σ from the set of lower-level worlds Ω to the higher-level worlds Ω' is a many-to-one function.

There is a connection between levels of grain and ontological levels: "if each coarsened partition of the underlying set Ω is re-interpreted as a set of higher-level worlds, then the given 'level of grain' will fit the formal definition of ontological levels" (List, 2017, p. 13). Although all systems of levels of grain are thus interpretable as systems of ontological levels, the converse isn't true: a system of ontological levels without a lowest level is not equivalent to any system of levels of grain. A further difference is that a system of ontological levels is a more general object than a system of levels of grain: higher-level worlds might be picked out by equivalence classes of lower-level worlds, but they needn't get *identified* with them and so they might still 'have their own spirit', or in non-metaphorical terms, these higher-level worlds might still have features of scientific or philosophical importance. Sometimes the language used is that the higher-level facts are something 'over and above' the physical facts (despite supervening on them).

Here we run into the difficult controversy about property identity, which cannot be resolved by appeal to supervenience (Horgan (1993)). For example, there is a set of physical worlds where my hand hits the face of someone who has in no way provoked me — and the equivalence class of the different ways I could do this would correspond to the higher-level worlds 'Katie does some morally reprehensible punching': but these worlds (and the properties instantiated in them) might have their 'own spirit'.

For instance, Moore (1903) and Hare (1952) agree that if my punching you is wrong in this world, then it is also wrong in a world identical to this one (identical with respect to say, the 'physical facts'). Twin Katie also should not punch people. As such, fixing the descriptive facts fixes the moral facts. But despite both subscribing to a supervenience thesis about the moral, namely that moral facts supervene on natural or empirical facts, Moore and Hare disagree about the nature of moral facts. According to Moore, moral properties have their own nature, whereas Hare believes 'wrongness' merely expresses our psychological attitude to punching (and includes a recommendation not to punch people). Thus, an A family of properties might supervene on a B family of properties, but this doesn't fix the *nature* of the A properties.

Explicitly distinguishing ontological levels from other systems of levels is helpful. This is because it makes clear that the metaphysical implications of 'levels talk' is a further project beyond considering how the different levels described by different sciences relate to one another. Even if we understand a system of levels of description or dynamics, how this relates to a system of ontological levels is still to be considered. Here, the relationships between different *systems* of levels comes in.

We can find a functor from a system of levels of description or dynamics to a system of ontological levels, but in general this functor will be forgetful in the category theoretic

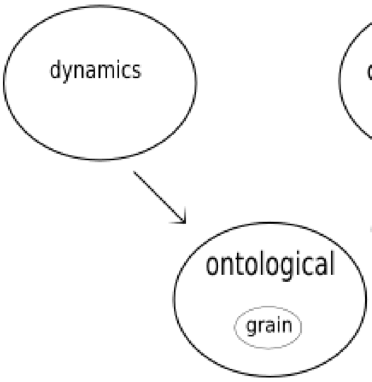

Figure 1.3: How different systems relate. Here the arrows represent forgetful functors. That levels of grain are a subset indicates that they are a special case of ontological levels.

sense: levels of description encode more information than ontological levels. As such, "different systems of levels of description involving different level-specific languages could induce structurally equivalent systems of ontological levels" (List, 2017, p. 16). Different languages can in principle be used to describe the same sets of level-specific worlds and level-specific properties, while describing them differently.

For example, levels of description have a minimally rich set of worlds Ω_L that settle everything that can be said in L. But there might be some 'descriptive fluff' in L (as is familiar from the literature on 'gauge' quantities in physics) and so we should commit to the system of ontological levels that corresponds to 'forgetting' this descriptive fluff.

1.3 Supervenience

In this Section, I discuss the notion of supervenience, since one central feature of List's framework is the supervenience mappings, σ. In Section 1.3.1, I explicate what supervenience is, and why it is such a central topic. I then argue, in Section 1.3.2, that the idea of global supervenience is plausible.

1.3.1 The formal idea

A set of properties, A supervenes on another set of properties, B, if there is no difference in A without some difference in B. But this can be an asymmetric relation: there could be some difference in B without any difference in A. Another way of putting this is that there are no two worlds which agree on all the B facts, but disagree about the A facts. Because of this asymmetry, the A properties are often called the 'higher-level properties'. Within the context of different scientific theories, a claim of supervenience would be: the higher-level trajectories/entities/properties described by that theory T_t

(‘t’ for ‘top’) supervene on the lower level T_b (‘b’ for ‘bottom’). That is, there are no two worlds which instantiate the same trajectory/entities/properties described by T_b but differ in their T_t trajectory/entities/properties.

The idea of supervenience is closely related to implicit definability in model theory. Once a certain realm of facts are fixed (i.e. once the B facts are fixed), that fixes the rest of the facts about that structure (i.e. the remaining, i.e. A facts).

Aside from List’s framework, supervenience is a central philosophical topic, because of its connection to the doctrine of physicalism: in a slogan, the thesis that all the facts are fixed by the physical facts. That is, all the higher-level facts (say, about our mental states) supervene on the physical facts: there are no two worlds which are identical with respect to their ‘physical facts’ but differ with respect to their higher-level facts (say, the ‘mental facts’). However, physicalism is a contested thesis[6] — and not one I will discuss here.

Next, I discuss why I think the assumption of supervenience in List’s framework is plausible.

1.3.2 The plausibility of supervenience

Earlier in Section 1.2 we saw that in List’s framework, the map σ expresses the idea that the higher-level supervenes on the lower level.[7] Why should we think that this is a plausible assumption?

The challenge is as follows. For level A to supervene on level B, for every distinct state or possibility described by A, there must be at least one distinct state described by B. That is, the map from states of B to states of A must not be one-many (hence the requirement that σ is a function). Furthermore, this function must be surjective: there must be no states of A ‘left out’ by the map from B to A.

A necessary (but not sufficient!) condition for A to supervene on B is that the number of states in B (that get mapped to A) must be at least as large of the number of states in A. (This condition does not suffice because even with more states in B than A, the map might go from one state of B to five states of A: if this were so then, as above, it would be possible to have difference in A without a difference in B).

Why think that this necessary condition will (generally) be fulfilled? Taking inspiration from Wilson (1985), one reason to think that there will be a physical description

[6] Not only is physicalism a contested thesis, but to echo Crane and Mellor (1990), Butterfield (2011a) and Dasgupta (2014) amongst others, care must be taken to render physicalism a coherent and substantive thesis. In particular, stating that the supervenience basis is the ‘physical facts’ is a fudge. Which physical facts? The facts described by our current most fundamental theory? Or of an imagined completed final theory? The danger is that the ‘physical facts’ will be just defined to be those facts which fix everything else; thus rendering physicalism trivial. See, e.g., (Butterfield, 2011a, §5.2.2) for further discussion.

[7] List is talking about the supervenience of possible worlds, so this is global supervenience - a weaker thesis than local supervenience, cf. Teller (1984).

corresponding to the higher-level description is the richness of mathematical functions. In order to secure supervenience, there needs to be at least one physical description of a given state of affairs. If not, we could end up with the 'gappy' situation of a 'something' at the A level without a corresponding 'something' at the B level: in this way, the facts at the A level would outstrip the facts at the B level. For example, the higher-level description could be the chime of a grandfather clock, the currents in a river or the splitting of stem cells. The reason to think that these processes will supervene on processes at the lower level (and indeed at the fundamental level, if there is one) is that there are just so many possible mathematical functions. For example, there is a function that describes the trajectory of the centre of mass of Trump's left hand, my cat Tibby and the football that Son used to score Tottenham's last goal.

This seem counter-intuitive; undergraduate physics courses teach us that physics can only deal with highly specialised situations (hence the joke: 'assume a spherical chicken'). Finding 'closed-form' equations and doing calculations is very difficult. Thus, writing down the evolution of the above complicated centre of mass seems well-nigh impossible. Or less flippantly, writing down a predicate such as 'unemployment in the Cambridge area' or 'inflation of 5 per cent' in the language of physics seems impossible.

Admittedly, it is unclear that we could write such a function describing Trump's hand, my cat and Son's football down. But nonetheless the richness of the real functions means that somewhere in the Platonic heavens this function exists. This is because the function need not be nice — like the type that we deal with in physics. But in full generality, a function is just a list of ordered pairs. Thus, the moral of Wilson's paper is that a function for the position and momenta of anything can be described by physics. (Of course, this way of putting the point assumes a classical world view. But there is no reason to think that this descriptive richness would go away when moving to the mathematics of quantum mechanics).

Does establishing (the plausibility of) the supervenience of the higher-level theories to the lower-level theories make reduction easily had? There are a range of formal connections between supervenience and reduction. For example, when the two theories under consideration are first-order theories and the other conditions for Beth's theorem apply are fulfilled, supervenience collapses into reduction.[8] But the assumptions required for Beth's theorem do not hold for many realistic cases of scientific theories, since our scientific theories are not first-order formal languages.

List (2017, p. 31) gives a 'combinatoric' argument as to why reduction (in a certain sense) can be elusive despite supervenience. The moral is: the number of possibilities at each level is vast, and the supervenience map needn't be nice — it could just be a list of elements of equivalence classes. This not only makes getting hold of σ daunting,

[8] The first philosophers to emphasise Beth's theorem as threatening such a collapse were Hellman and Thompson (1976), and Butterfield (2011a) discusses this topic in detail.

there are in-principle reasons to think it is sometimes impossible.

But fascinating though these formal connections between supervenience and reduction are, they will not be the focus of this thesis. Instead, my focus is on particular case studies of reduction — and any misgivings about assumptions of supervenience making reduction 'too easy' are quelled by considering the difficult details and controversies that surround any putative case of reduction.

1.4 Multiple realisability

We have already encountered multiple realisability within List's framework, where a higher-level world w' has multiple lower-level worlds w that realise it: in other words, σ is not injective. Multiple realisability is also discussed in terms in *properties.*[9] A (homogeneous) higher-level property is multiply realized when it is realized by many distinct (heterogeneous) lower-level properties. As such, the map from the lower-level properties to the higher-level properties is many-to-one. A common example in the literature is jade. Being jade is a higher-level mineralogical property that is multiply realized, as it is realized by two chemical properties: 'being jadeite': $\mathrm{NaAlSi_2O_6}$, and 'being nephrite': $\mathrm{Ca_2(Mg, Fe)_5Si_8O_{22}(OH)_2}$. Another popular (hypothetical) example: 'pain' is realised by 'C-fiber firing' in humans but 'D-fiber firing' in octopuses.

In this Section, I review why various authors consider multiple realisability to be a problem for reduction — and I offer several reasons to think that multiple realisability (MR) need not block reduction. In fact, I will go further: I will argue that MR is very widespread. Indeed, Mellor (1978) goes as far as to say that any property is multiply realised by its instances. Yet this threatens to make MR trivial, which leads Shapiro (2000) to offer an account of 'substantive MR'. Nonetheless, in Section 1.4.2, I will argue that reducing the higher level T_t to the lower-level T_b makes the seemingly surprising (or substantive) cases, unsurprising or non-substantive.

1.4.1 Multiple realisability: a problem for reduction?

In this Section, I outline the examples from the literature which argue that MR of a higher-level state (or 'kind') M by multiple lower-level states (or 'kinds') $(P_1 \vee P_2 \vee P_3)$ is a problem for reduction. This is known as the multiple realisability argument (MRA). Different authors think that MRA causes problems for different reasons. For example, we will see in Section 1.4.1.1, that for Fodor the lower-level disjunction threatens the law-like nature of statements involving that disjunction. For Putnam, in Section 1.4.1.2, the disjunctive lower level cannot provide an *explanation* unlike the higher level. Whilst

[9] If a possible world is defined to be the histories of all the properties of all the objects in that world, then the world and property view of multiple realisability come together.

Fodor emphasises laws and Putnam emphasises explanation, the problem for reduction from MR is nonetheless treated as a single argument in the literature, which I will rebut, following Sober (1999).

1.4.1.1 Fodor's horror

Fodor (1968, 1975) argues that multiply realisability (MR) blocks reduction. In his well-known diagram (cf. Figure 1.4), the kinds of the special science S_1, S_2 are multiply realised by physical kinds P_1, P_2, P_3. This has the consequence that the higher-level law $S_1 \rightarrow S_2$ is translated as $(P_1 \vee P_2 \vee P_3) \rightarrow (P_1^* \vee P_2^* \vee P_3^*)$.[10] For Fodor, a higher-level theory is reduced to a lower-level theory if the laws of T_t are explained by the laws of T_b. Not only must we explain, using the resources of the lower level, why the laws of T_t are true, but also why they are laws. In order that the generalisations of T_t are bona fide laws, they must be derived solely from law-like statements. But Fodor claims that disjunctions cannot feature in laws, and so MR prevents reduction.

To link this to the earlier discussion of levels of description: even if the propositional content of a higher-level explanation $[\phi]$ can be expressed in the lower-level language L_B, the 'peppering' of the explanation with the word 'or' prevents this from qualifying as reduction in Fodor's eyes.

Thus, Fodor is committed to (at least) two assumptions: (1) laws must be reduced to laws; (2) disjunctions cannot figure in laws. Sober (1999) argues that Fodor is motivated by considering natural kinds — and Goodman (1954)'s example of grue and bleen. Allowing disjunctions into laws comes at a price that Fodor is not willing to pay: one can no longer 'read off' natural kind predicates from a law statement.

But I reject Fodor's motivation, since interpreting the metaphysical implications of our scientific theories is not a matter of reading off natural kinds predicates. Thus, I believe the MRA does not prevent reduction, because there is no reason to think disjunctions cannot figure in laws, as Sober (1999) argues.

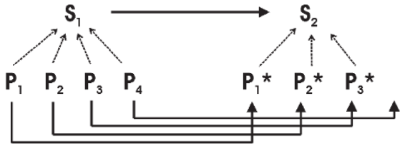

Figure 1.4: Fodor (1968)'s multiple realisability diagram.

[10] There's a certain irony that he sees this as presenting a problem for reduction. In the physicist's sense of reduction if you can demonstrate such a diagram then you've got reduction sorted!

1.4.1.2 Putnam's peg

Putnam (1967, 1980) uses the MRA to block reduction by focussing on the explanation of singular occurrences. His famous example is that of a square peg not fitting into a round hole. There are many different microphysical configurations compatible with a certain dimension of the peg and hole: these properties (the hole, the side length of the peg) is realised by a multiplicity of pegs of different colours, materials, temperatures and weights (to name just a few 'microphysical' details). But these details are irrelevant to the geometrical explanation of why the peg doesn't fit through the hole: its sidelength is too large.

Putnam claims that whilst we could give a micro-detailed description of the peg, there would be too many extraneous details for this description to be an explanation of why the peg doesn't fit. And he says: because the lower level (the microphysical level) cannot explain the higher level, the higher level is not reducible to the lower level.

Here, once again, I think Sober (1999) offers a convincing reply. The lower-level description might give you more details than you want to hear, and you might say its not the *best* explanation; but it is nonetheless *an* explanation. Generality (i.e. breadth) and depth (i.e. being detailed) are competing virtues in an explanation. Which you prefer is not an objective matter — and so anti-reductionists are mistaken to think breadth trumps depth *tout court*, just as reductionists (or 'eliminativists') are mistaken to think that more details, and so depth, trump generality *tout court*. Whether depth or breadth is better depends on the context. Sometimes more details cloud the relevant facts, and sometimes the details lead to more precision. It depends on the situation and what you care about.[11]

To sum up: Putnam's argument requires that (1) explanation is a core component of reduction and (2) a given phenomenon can only be explained at one level; there cannot be both microscopic and macroscopic explanations. But I reject both (1) and (2).

1.4.2 Multiple realisability is widespread — and non-substantive?

I have argued that MR is not worrying for reduction. But the debate surrounding the MRA implicitly assumes that multiple realisability is a substantive or special phenomenon. Yet in List's framework, multiple realisability is widespread. In this Section I outline why, in general, MR is widespread, and consequently why it is tempting to want to offer a criterion to distinguish between substantive and non-substantive cases in order to retain the mystery of MR as a special phenomenon. But I will argue against this strategy and I claim that the project of reduction between the

[11] As (Sober, 1999, p. 549) notes, Putnam would not accept this reply, since he believes that the "goodness" of an explanation is not a "subjective" matter. Context-dependence needn't imply subjectivity, but nonetheless, this reply is unlikely to convince Putnam.

two levels in question, will render seemingly substantive, or in more psychological terms: *surprising*, cases of MR akin to the widespread non-substantive cases.

1.4.2.1 Substantive and non-substantive MR

It is easy to find examples of multiple realisability. The property 'blue' is multiply realized by aquamarine and cobalt. We can "simply point out that we employ different standards of accuracy at different levels and so it is unsurprising that we see heterogeneous properties on one level and homogeneity on another" (Mainwood, 2006, p. 123). This is clearly an example of List's levels of grain.

Furthermore, even a maximally specific shade of blue can be multiply realized by a painting or a computer pixel. Thus, the determinate-determinable relation is another sense in which MR can be unsurprising. The property blue is multiply realised by its different shades.

But one might worry that this *trivialises* the idea of MR: "the lower-level realizer properties are distinct only due to differences *manifestly irrelevant* to the higher-level properties they realise" (Mainwood, 2006, p. 123) (emphasis added). Pain can be realised by neurons and by neurons stained purple: the staining is manifestly irrelevant.

In response to this abundance of trivial MR, Shapiro (2000) aims to articulate a definition of *substantive* MR which hones off the irrelevant ways in which a property might be multiply realised. Shapiro's criterion for 'substantive' cases of multiple realisability is: MR is non-trivial only when realisers differ in an aspect 'causally relevant'. For instance, pain can be realized both by neurons and by neurons which have been stained purple, but this is mere trivial, non-substantive, or unsurprising MR since purple staining is causally irrelevant. The aim of his account is to give a criterion for multiple realisation that captures the interesting or substantive cases, whilst ruling out the manifestly irrelevant cases such as purple staining and differing standards of accuracy.

However, I do not think Shapiro's criterion succeeds in capturing the right cases. Pain is meant to be an exemplar case of MR and so — if anything does — pain should surely count as 'substantive'. Yet, pain being realized by ordinary neurons and silicon-based lifeforms would also count as trivial, as the causally relevant aspect of electrical activity is present in both.

Furthermore, if causation is understood as 'difference-making' (cf. Woodward (2005)), then the lower-level details are by definition causally irrelevant. Whichever ways the lower-level realisers differ (and however surprising they might be), by definition these differences won't matter for instantiating —i.e. causing the existence of— the higher-level property.

Thus I am skeptical that Shapiro's criterion successfully distinguishes substantive from non-substantive cases. But furthermore, I think that sorting 'trivial' from 'substantive' cases of MR is unproductive, because it just depends on how well-understood the connection between levels are — i.e. whether the reduction has been carried out yet or not.

1.4.2.2 MR and reduction

Working out which lower-level differences are irrelevant for the higher level is part of the reductive project. For example, we are unsurprised that neurons and neurons stained purple form an equivalence class and the difference — i.e. the staining — is irrelevant; no one ever thought that staining would be crucial. But the case of carbon and silicon-based life differs. Electrical signals are the crucial feature or commonality across carbon and silicon-based life: but it is only once we understand the centrality and importance of electrical signals that the differences becomes manifestly irrelevant. Once we understand the relationship between the higher-level state and its lower level realisers — e.g. once we understand the relationship between electrical signals and pain — we can often see why the lower-level details didn't matter. In connecting two levels of description as is done in reduction, the irrelevance is made manifest. If we understand why the two lower-level states s_1 and s_2 realise the same higher-level state s', then we can understand why the higher level is impervious to change between s_1 and s_2.

Whether an example of MR appears substantive seems to hinge on how surprising we find it. In non-substantive cases, the differences between realisers is manifestly irrelevant. But there are cases where it is not obvious that the differences shouldn't matter. As such, that the lower-level states s_1 and s_2 form an equivalence class, may be *surprising*.

This is what Papineau (2010) emphasises — if there were a higher-level law that 'plastics dissolve in lakes' but this law was MR by a variety of physical mechanisms with nothing in common (in one lake an acid dissolves the plastic, another is so hot it melts etc.), then the higher-level regularity is surprising.[12] So the lower-level states differ in ways that are *not* manifestly irrelevant for the higher level. Thus, it often might not be obvious — i.e. manifest — why the lower-level states form an equivalence class: i.e. why the differences are irrelevant.

Within List's framework, the lower-level states $s_1, s_2 ... s_n \in \mathbb{S}$ form an equivalence class that gets mapped by σ to the higher-level state $s' \in \mathbb{S}'$. What do $s_1, s_2 ... s_n$ have in common? According to the formal framework, they need not have anything intuitive or 'natural' in common (such as being close to one another on the colour wheel).

[12] Papineau goes on to claim that as a consequence the higher-level regularity cannot be a law, but I resist this move since whether a regularity counts as a law does not depend on issues of reduction as laws don't inherit their status from the lower level.

Mathematically speaking, there are no rules about what forms an equivalence class: each class could be a random list of elements. If the realisers seemingly have little in common, it will certainly be surprising that they realise the same higher-level state or world. Indeed, were Papineau's examples true, they would be surprising.

Papineau claims that there must be some unifying feature of the lower-level realisers. Establishing how the higher-level world arises from the lower-level will mean that the equivalence class will no longer look like a random list. And according my account, this is done through 'reduction-as-construction'. Thus, through reduction we will remove the surprise, because we will understand why particular lower-level states form an equivalence class. Thus, the MR will seem non-substantive.

But, by way of criticising Papineau, note: whether the equivalence class of lower-level realisers is 'unified' is vague. Papineau portrays the anti-reductionists as claiming that there is *nothing* in common at the lower-level, but this is uncharitable. There is a limit to how different the realisers are: after all, they are described by the same theory T_b. Furthermore, how 'similar' two lower-level states, or worlds, count as is notoriously vague. There are so many — infinitely many! — ways to make equivalence classes out of the lower-level states. Aside from brute enumeration of random elements, one could form equivalence classes of states depending on the value of one independent variable such as position, whilst abstracting away from all values of momenta. Alternatively, the states could be members of the same equivalence class if they have the same average over several variables. There so many possible partitions and it unclear that one partition unifies the lower-level states more than another. Indeed — and some may think this a concession to the anti-reductionist— note that: in List's framework, it could be that the equivalence classes can only be 'unified' (for want of a better word), by using the higher-level language.

To end with more wisdom from Sober: "This is the kernel of truth in the MRA: the higher-level sciences 'abstract away' from the details that make for differences among the micro-realisations that a given higher-level property possess" (Sober, 1999, p. 560). This phenomenon is widespread and nicely encoded in List's framework.

1.5 The reduction-as-construction account

Throughout the preceding Sections I have considered the literature surrounding accounts of reduction such as Putnam's explanation and Fodor's laws. Now, I outline my own views on reduction. In Section 1.5.1, I answer the question: reduction between what? The aim of reduction is to capture the behaviour described by T_t in terms of T_b. But because one theory can describe a diverse range of behaviour, it might be that it is between particular *models* of each theory that the reduction relation holds. As such reduction might be said to be local.

In Section 1.5.2, I outline the core of my account: reduction-as-construction. I stipulate that if the equations or quantities described by T_t can be constructed from the lower-level theory T_b, then T_t is reduced to T_b. This focus on the mathematical relations, rather than logical relations fits with a theme of this thesis: I am concerned with reduction in practice rather than in principle (cf. Section 1.8). In Section 1.5.3, I suggest that whether the lower-level theory must capture the equations *or* the quantities lines up with whether the higher-level theory is dynamical in a certain sense. The special sciences are not dynamical theories in the way familiar from physics — and this leads to an important distinction. Sometimes we are concerned with reducing an older theory to its successor, but —as is familiar from the discussion of the special sciences— sometimes we are concerned with reducing a higher-level, or macroscopic theory to a lower-level, or microscopic, theory. In Section 1.5.4, I spell out this distinction, and suggest it is a difference in degree, rather than kind.

Having spelt out my view of 'reduction as construction', I will then suggest, in Section 1.6, that functionalism can be useful for securing reduction. But I will reject Kim's account — and his eliminativist aim. Indeed, in Section 1.7, I will argue that the aim of reduction is to vindicate, not eliminate the higher-level. This will lead to spelling out my metaphysical position — and discussing how my account of reduction is compatible with emergence. Finally, in Section 1.8, I discuss the distinction between reduction in principle and practice.

1.5.1 Reduction between what?

Different accounts hold that the asymmetric reduction relation holds between different relata: properties, theories, models...to name but a few. Indeed different accounts' requirements on reduction only make sense between certain relata. For instance, it is unclear what it means for one theory to be a limit of another. Limiting relations are more usually defined to hold between mathematical functions. Thus, it is more precise to say that one quantity (represented by a mathematical function) reduces to another: $\lim_{n\to\infty} Q(n) \to P(n)$. I am going to stipulate that reduction holds between theories, but sometimes this should be understood as a local matter — sometimes it is better to say that one model reduces to another and so reduction is a local affair.

First I want to emphasise, following other authors, that it is "the behaviour characteristic of the system [that] is the focus of reduction" (Rueger, 2006, p. 343). Similarly, Rosaler (2017, p. 4): "*Fundamentally*, the concept of reduction that we investigate here is about showing that all real behaviours that can be accurately modelled in one theory can be modelled at least as accurately in another. Taking limits and deriving one set of laws from another may turn out to be useful *strategies* toward this goal, but neither requirement is regarded from the outset as a *sine qua non* of reduction".

Theories can describe a wide range of behaviours. For example, the regular swing of a single pendulum and the chaotic motion of a double rod pendulum are both described by the same theory: classical mechanics, T_{CM}. Especially clear examples of different behaviour between two histories of the same theory can be found by considering time-reversed trajectories. Releasing the molecules from a balloon will lead to these molecules exploring the entire room. But the time-reverse of this trajectory represents all the gas molecules converging together and all entering the balloon. Both are possible histories according to the lower-level laws, but they display very different behaviour and so might be described by different higher-level theories. (This example will be central in chapter 3).

As an example, take Newton's third law, $F = ma$. This law applies to wide range of different systems and there are many different initial conditions which generate a range of different possible histories, which are the possible worlds according to T_{CM}. Yet these histories can be qualitatively very different — think of a collection of particles with random trajectories or perfectly aligned trajectories — depending on the initial conditions. Thus, there is a range of behaviours encompassed by the one theory. Indeed, the same equation can govern a wide range of different behaviours.

Because the histories, or possible worlds, describable by one theory are diverse, they might not all give rise to higher-level worlds. For instance, not all physically possible worlds will give rise to biological worlds: the conditions could have been not 'just right' to give rise to life. Less speculatively, there are quantum mechanically possible worlds that do not give rise to classically possible worlds: those worlds where the environment is such that decoherence does not happen (Schlosshauer, 2007). Or less controversially, there are CM histories (i.e. the 'going-back into the balloon' type case) which do not give rise to thermodynamically possible worlds. As such, the lower level can countenance possibilities that do not give rise to the higher-level regularities.

Thus, as seen in Figure 1.5, and as described earlier, the (pre-image of the) set of possible histories (or worlds) according to T_t may only be a subset of the possible histories according to T_b. To sum up: because we are concerned with modelling 'behaviour', reduction is a local affair.[13] The reduction relation might not be global but instead different detailed stories might hold for different systems, or models.

1.5.2 Reduction-as-construction

'Reduction' between two theories occurs when the behaviour of phenomena described by T_t can be captured by the lower-level theory T_b. Usually, in a highly mathematised

[13] This local nature of behaviour, rather than MR, is the reason that reduction is sometimes local, rather than global.

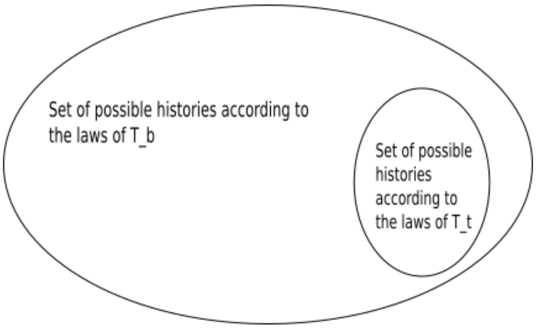
Set of possible histories according to the laws of T_b
Set of possible histories according to the laws of T_t

Figure 1.5: The pre-image of the higher-level possible histories might only be a subset of the lower-level possible histories.

science such as physics, this will involve *constructing* the key equations or quantities described by T_t in the framework of the lower-level theory T_b. *Construction* is conceived informally: the physicist can use whatever is required to get the job done. The idea is: start with the expressions or models of both T_t and T_b, and then perform whatever conceptual or mathematical manipulations are required to take you from one model to another. Admittedly, this is not a detailed specification of 'construction' — but instead of a weakness of my account, I see this as a strength: physicists should use whichever techniques they require, without being hamstrung by philosophers' restrictions. Typically, construction might involve defining new variables (perhaps by summing or performing other 'irreversible' mathematical operations) and a whole range of approximations, idealisations — and whatever mathematical or computational tricks the physicist can lay her hands on.[14]

This is not intended as an alternative account to a Nagelian account. Rather the focus is different. Whereas Nagel focuses on logical ideas: T_t is reduced to T_b if T_t (or a close cousin) can be deduced from T_b augmented by appropriate definitions. Reduction-as-construction focuses on the mathematical features of reduction: T_t is reduced to T_b if the key features (i.e. equations of motions and key quantities) of T_t can be constructed (for a particular domain or for a particular level of accuracy) using whatever mathematical resources are available from the key features of T_b.

Here are two illustrative examples.

1. The equations of Newtonian mechanics can be constructed (or recovered) from special relativity by taking the $\frac{v}{c} \to 0$ limit (Batterman, 1995).

2. The irreversible equations of statistical mechanics can be constructed from the

[14] One might worry that such laissez-faire attitude to approximation and idealisations leads to dialectical issues. In particular, these approximations and idealisation are exactly the type of issue that anti-reductionists such as Batterman think prevent reduction. Is this account of reduction worthy of the name? Would an anti-reductionist oppose this account? I am happy if they do not: for if that is so, I will have succeeded in my aim to capture what is right about both the reductionist and anti-reductionist camps. I am grateful to Alex Franklin for this comment.

Liouvillean dynamics by coarse-graining the probability distribution and making several assumptions (such has an initial state condition and the Markovian approximation), as I will extensively discuss in Chapter 3.

1.5.3 Dynamical vs. 'non-dynamical'

I claim that to capture the behaviour described by T_t in terms of T_b, the equations or quantities of T_t must be constructed from the equations or quantities of T_b. In this Section I explain why I used the phrase 'equations or quantities'. Equations and quantities are, of course, not unconnected: the dynamical equations describe how various quantities evolve over time. I first consider how 'meshing dynamics' naturally fit with the dynamical theories that physics is concerned with. But I then discuss how 'dynamics' in this sense are not always central, by considering special sciences and Fodor's diagram.

Dynamical theories, in the sense familiar from physics, are considered by an account given by Rosaler (2015, 2017), which is version of the levels of dynamics discussed in Section 1.2.3. The important difference from the initial discussion of levels of dynamics is that here we start with the higher-level dynamics, which have already been specified independently by T_t. We can then compare an original higher-level dynamical trajectory with a trajectory induced — as we saw in the toy example of Section 1.2.3 — by the coarse-graining.

A dynamical theory specifies a state-space (which represents different possible states of the system K), and a dynamics D (which specify K's trajectory through this space). If there are two different dynamical theories T_b and T_t describing K, there will be a state-space and a dynamics for each theory: $\mathbb{S}_h$ and D_h , $\mathbb{S}_l$ and D_l respectively. The solid lines in figure 1.6 show the trajectories in each state-space representing K's dynamical evolution.

Rosaler describes the meshing dynamics as follows. The map B (or in our earlier notation σ) takes states in $\mathbb{S}_l$ to states in $\mathbb{S}_h$, cf. Figure 1.6. The dashed line $B(D_l(x_l, t))$ is the induced dynamics, or the higher-level image of the lower-level dynamical trajectory $D_l(x_l, t))$. The solid line $D_h(B(x_l, t))$ is the dynamical evolution of K according to the high-level dynamics D_h. Here the map B commutes with the dynamical evolution of states in S_l: we have the meshing situation of Section 1.2.3.

Now we can compare the induced macrodynamics with the original higher-level dynamics — and we see that they match, but only approximately. That is, the image is not identical: the two theories might only agree within a certain margin of error, δ. Furthermore, the dynamics and coarse-graining may only commute for certain domains $d_l \subset S_l$. That the errors do not accumulate to render the higher-level description useless is part of Duhem's principle of stability (cf. Fletcher (2017)).

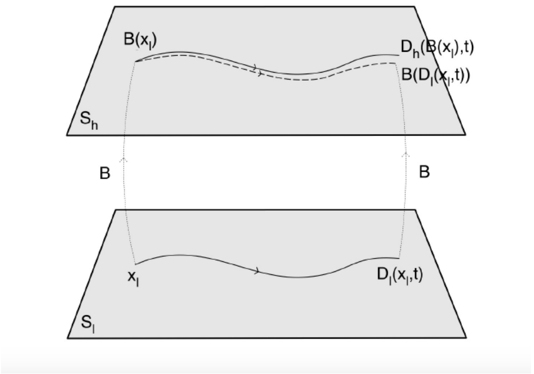

Figure 1.6: Diagram from Rosaler (2015)

The meshing idea is illuminating — but not always central. It is only appropriate for certain theories familiar from physics: that is, the theory specifies a state space of physical possibilities for the system and then the dynamics specifies how a system spontaneously moves from one point to another in that state space. Then the trajectories through that space state can be compared in the above manner. But many sciences (for example, arguably organic chemistry) does not have dynamics in the way that physics generically does. Likewise, as I discuss in chapter 2, thermodynamics is an unusual physical theory in this way: it does not discuss spontaneous trajectories. Thus, it is unclear that Rosaler's account is applicable here.

Of course, there's an 'in principle' sense in which all systems 'have a dynamics', in that their state changes over time! But this not always a detailed or central part of the science describing that system. The main concern is not recovering different equations of motion.

This is made especially clear in Fodor's famous diagram 1.4, displayed earlier in Section 1.4.1.1, about the relationship between physical kinds and special science kinds. Here Fodor has just assumed that the 'meshing' situation obtains — a substantive assumption. (See Papineau (2010) and Butterfield (2012) for evolutionary reasons to think Fodor's hope is not foolish). But here the key point is that Fodor is not concerned with the left to right arrows— how the *dynamics* at the different levels relate to one another. Instead, the concern is how the *quantities* P_1 and S_1 relate to one another: the up-down arrows.

If the two theories under consideration are not dynamical theories as in physics, then the concern shifts to whether the quantities of T_t used to capture the behaviour of the phenomena in question can be constructed from T_b. (To flag a connection to Section 1.6: I have stipulated that 'capturing behaviour' is crucial to reduction. Behaviour will of course be centre-stage when considering equations of motion or dynamics. But in these 'less dynamical cases' where temporal change is implicit, functionalism may be helpful for keeping the focus squarely on behaviour.)

1.5.4 Two kinds of reduction: vertical and horizontal

In this Section, I will distinguish two types of reduction. In one case, T_b is the 'better' theory and T_t is the 'tainted' theory — T_b offers an improvement over T_t. I will label this 'horizontal reduction'. This horizontal reduction is often exemplified when an older theory is reduced to its successor. In certain limits, the old theory's description of a given system's behaviour is very similar to, i.e. *approximates*, the new theory's description. In the second case, T_b is the 'bottom' theory and T_t the 'top' theory — I will label these 'vertical reduction'. In these latter cases, the different levels describe different subject matters. For example, T_t might describe the macrolevel, and T_b might describe the microscopic realm. The crucial distinction is between reduction as improvement (horizontal) and reduction between different subject matters (vertical). An alternative labelling would be 'diachronic' and 'synchronic' reduction, but these labels elicit temporal, or historical, considerations which I do not consider.

These two types of reduction have different philosophical consequences. Whether there are examples of horizontal reduction, and how extensive they are is important for the scientific realism debate. For if we can construct our old theory from our new theory we can see how the old theory was successful, despite its falsity — and we can articulate the part of it which is 'approximately' true and will be preserved over theory change.

The presence (or lack of) vertical reduction has historically been taken to have metaphysical implications. Carnap and the Vienna Circle, for instance, were interested in establishing the unity of science. If there were no vertical reductions at all, one might wonder in what sense physics is primary (and so wonder about whether physicalism is true). Are the imperialistic overtures of physics justified? In particular, the existence of vertical reduction is important for the debate about the autonomy of the special sciences — and the plausibility of non-reductive physicalism.

In Section 1.5.4.1, I argue that approximation is crucial to horizontal reduction and in Section 1.5.4.2, I argue that abstraction is crucial to vertical reduction. Then, in Section 1.5.4.3, I question how strong the distinction between approximation and abstraction is, and demonstrate how the two can come together in particular cases. This emphasises that the difference between horizontal and vertical reduction is one of degree, not kind. My aim in drawing this distinction is to emphasis that T_b is not always better. This will not only be important for Section 1.7's metaphysical considerations, but also in Section 1.5.4.4 to give a reply to Putnam.

1.5.4.1 Horizontal reduction: reduction as 'improvement'

In the case of horizontal reduction, T_t is the *tainted* theory. That is, it is someway inferior to T_b — most likely, there will be situations where it is less empirically successful. Hence, T_b is the successor theory to T_t. But, in order to have been accepted in the first place, T_t will of course be empirically successful, but just to a lesser extent than T_b. Frequently, T_b reveals that domain of T_t was limited — it was successful for certain problems, but not for all. And the successor theory T_b can explained *why* T_t was successful in these domains. Often it is because T_t *approximates* T_b within certain confines. (I agree with Rosaler (2017) that what counts as a successful approximation depends on the situation, and so is an empirical, and local, matter).

Until this point, I've assumed that reduction is a two-place relation. But here we can see that centrality of approximation requires that it is a really a three-place relation. I endorse Rosaler's claim that empirical information about the world is often required into work out whether one theory reduces to another. In this sense, formal approaches to reduction that treat reduction as a two-place relation between theories or models which can be analysed using purely mathematical or logical resources miss out on something important. That important thing is that approximations are often used, and the world (i.e. empirical input) tells us which approximations we can 'get away with'. (Mathematically we can show how one thing approximates something else in a limit, but if the circumstances of the limit are not even approximately realised in the world, then this is not very enlightening). Thus, reduction is a three-place relation between some set of systems (i.e. patches of the world) and two descriptions of that patch. Rosaler (2017, p. 2): "Domain subsumption is taken to rest not only on an abstract analysis of logical or mathematical relations between these representations, but also on further empirical input concerning where they succeed at describing real physical behaviours". One obvious point is that we don't want to worry about recovering the stuff that T_t gets wrong! Thus, we must also specify the domain of the theory: (i) the set of systems well-described by models of that theory, and (ii) the set of circumstances under which those models succeed (i.e. the timescales and level of accuracy).

I agree that reduction involves tangling with the messy details of where our theories are successful; this is because I think reductions often involve approximations, and whether and how these succeed is always an empirical matter. Rosaler gives the example of Ehrenfest's theorem, which shows the dovetailing of the mathematical structures of QM and CM. But this holds over timescales over where the ensemble spreading in the quantum model can be ignored. Thus, we need empirical input about the timescales of which classical model succeeds at tracking the alpha particles behaviour. Another example: as we will see in Chapter 3, the recurrence timescale will be important for thermal physics.

Hence for horizontal reduction, the successor theory T_b's domain will subsume the domain of T_t, because —to a certain degree of accuracy δ and within a certain domain of applicability— T_t and T_b give the same answers.

One key part of horizontal reduction is that the T_b is the successor theory, and so T_t might be thought to be eliminated or replaced. But the domain of applicability point above explains why the old theory might still be used in scientific practice. If the problem, such as taking astronauts to the moon, is within the domain of applicability and the description given in T_t is accurate *enough*, then it might be pragmatically well-advised to use the old theory. For example, using Newtonian gravity rather than General relativity to calculate how to take the astronauts to the moon is much more tractable.

Thus, it is more convenient to apply Newtonian mechanics rather than general relativity for calculating rocket trajectories, or NM rather than QM for baseball trajectories. But the idea is that rocket launches and baseball trajectories fall within the domain of applicability of the QM and GR — and furthermore, their domains subsume NM.

To sum up: A model of T_t is reduced to a model of T_b if the T_t model describes the same behaviour as the T_b model — to a certain degree of approximation δ and for a certain class of systems. Limits, such as $\frac{v}{c} \to 0$ and $\hbar \to 0$, can be useful for explaining why T_t approximates T_b in these domains.

Before describing vertical reduction, we should note the latitude in the term 'domain of applicability'. The requirement is appli*ability* so it just has to be that T_t *could* be applied, but doesn't have to be that *in fact* for all practical purposes it does get applied. Thus, there are three different precisifications, which I list in descending order of strength.

A given system K, previously successfully described by T_t, falls within the domain of applicability of T_b if:

1. There is a description of K within T_b which is in fact used in practice.
2. There is a description of K within T_b which *could* be used in practice:
    a) but sometimes isn't, due to tractability considerations.
    b) but sometimes isn't, because T_t offers a 'better explanation'.
3. There is —*in principle*— a description of K within T_b.

1.5.4.2 Vertical reduction: reduction between different scales

Sometimes different scientific theories seem to be talking about different things from one another; platypuses never come up in undergraduate quantum mechanics courses, for instance. As such, it seems that biology and physics have different subject matters.

The term 'subject matter' has a heuristic use – familiar from the retort 'but you are changing the subject', i.e. the concern or focus of a dispute. When I am altering an examination syllabus in order to help or hinder the average student's chances, my concern is not about how any individual student, with their various idiosyncrasies, will do: the concern, and so subject matter, is different – how hard an exam it is.

One way to see that two levels could have different subject matters, but nonetheless be connected, is if abstraction is involved. Abstraction involves purposefully leaving out details. The formal counterpart of this idea is 'forming an equivalence class', where the objects in that class can differ in certain respects but are identical with respect to a given attribute. That is, they can differ in all respects except the given attribute. To take an example from Frege (1968), the set of straight lines can be partitioned into equivalence classes according to which lines are parallel to one another; we abstract away from the respects in which the lines differ, with the exception of their directions.

Taking equivalence classes leads to a coarser description of the set of worlds, as is familiar from List's formal framework. Because you have purposefully thrown away details about the lower-level of description, these details are not part of the subject matter of the higher level. As such, the subject matter can be more or less fine-grained. Thus, a different level of description is one that discusses a different subject matter: this could be a matter of grain, or more generally, the subject matters might cross cut one another.

Different levels of description, or subject matters, are often described by different scientific theories: as with the platypus and the electron. And these different scientific theories uses different languages. We needn't look as far apart as the platypus and the electron: fluid mechanics talks of viscosity, whereas statistical mechanics talks of entropy, and chemistry of enthalpy. Another way of putting this point is that different scientific theories (and indeed levels more generally) talk of different variables.

In cases of horizontal reduction, the two theories are rivals, competing to describe the same phenomena and answer the same questions about the world. We don't need to talk about reduction at all in order to evaluate which is the more accurate theory. The old theory — where it was successful — often approximates the new, *more accurate* theory. In contrast, in cases of vertical reduction, at least prima facie, the two levels are not competing to describe the same patch of the world: the top theory might be talking about macroscopic phenomena such as crowd behaviour, whereas the lower level might be concerned with the behaviour of individuals. Furthermore, you can't (necessarily) predict herd behaviour from looking at the behaviour of individuals. The reason that studying lone individuals is unlikely to lead you to discover herd behaviour, or studying individual hydrogen and oxygen molecules is unlikely to lead you to discover the properties of water is the obvious point that when we are considering the many not the few, interactions are important.

As we increase the number of particles from 1 to infinity, there is no obvious qualitative change in behaviour from the CM perspective: we still only have two-body interactions. There are no new fundamental forces that appear at a certain number — but there are, nonetheless, new bulk properties described by higher-level theory. The higher-level of description— or macroscale subject matter— may discuss these bulk properties (such as the clustering of a crowd, or the polar bonds in H_2O) and abstract away from the microvariables.

Indeed, the variables of the higher-level L' in certain cases may dynamically decouple from the variables of the lower level L. As we saw with List's levels of dynamics, if the two levels 'mesh', then describing the evolution, that is the dynamics, of the higher-level state s' does not require knowing the evolution of the lower-level state s (and we will see a vivid example of this in Chapter 3). In fact, this 'dynamical decoupling' or autonomy is important for L' to be a useful—rather than gerrymandered— level of description. That this often happens is part of the contingencies (or good fortune) of our particular world — and I believe provides the answer to Loewer (2009)'s question: 'why is there anything other than physics?'

Because, in cases of vertical reduction, T_t and T_b have different subject matters, they are not, prima facie, competing to describe the same patch of the world. What does it mean for one theory to be more accurate than another theory? CM is more detailed than SM so you might think that therefore it is more accurate. Yet this impulse would suggest that the level of physics is always more accurate than the psychological level. But frequently this claim can't be adjudicated. This is because in order to adjudicate the comparative predictive success of two different theories, they need to be able to answer the same question (e.g. 'at what velocity will the ball hit the window?').

In order to compare the two theory's predictions, one must be 'translated' into the other. (Of course whether this 'construction' or 'translation' is deserving of the name - or is sufficient for - reduction is disputed). Once this is done, we can then ask about accuracy. First I discuss this translation, and then accuracy.

A quantity such as 'temperature' or 'pain' is not amongst the predicates of the lower-level language, say of CM or QM. The subject matter of SM includes describing the behaviour of quantities such as temperature. Such quantities are not outside the domain of CM *as such*, but CM does not readily give a more *accurate* description of these quantities: the only way for CM to answer the SM questions is to construct the SM equations from CM.

There are two ways of seeing this. Firstly, formally we can think of a 'translation' in the following sense. A sentence S at the higher-level has propositional content, ϕ. This is the function from the set of possible worlds w to the worlds where that sentence is true. Secondly, and more informally, the higher-level variables can be 'constructed' from the lower-level theory. This second way is especially relevant when the two levels describe different scales — the variables of the higher level can be constructed from the lower level. In certain cases, the higher-level variables will be collective variables.

Only once we have an image at the lower level of the higher-level description of T_t, does it make sense to claim that they are both describing the same system K and to ask which theory is more accurate. But we might think that there is a limit to how different the constructed, or translated, description will be from the original higher-level description, especially if the original higher-level, or macroscopic, variables were used to 'guide' the constructed variables. Nonetheless, whilst the bulk of the image will coincide with the original higher-level description (as that is what it has been engineered to do), there might be details at the fringes which differ. In particular, the lower-level image might reveal circumstances in which T_t is not accurate — and so limit its domain of applicability. This suggests a blurring of vertical and horizontal reduction, which I now consider.

1.5.4.3 Distinguishing approximation from abstraction

In abstraction, there is no aim to remove the abstraction in order to get a more accurate representation. This is unlike idealisation where de-idealising (if possible) would lead to a more accurate representation. Approximation is similar to idealisation in that there is an element of falsity in the description - and furthermore, one might want to remove this unrepresentative aspect (for certain purposes).[15] Whether an approximation (such as replacing $\sin\theta$ by θ) is a good one, depends on the system at hand: replacing $\sin\theta$ by θ is good for small θ — and what counts as small depends on the situation. As such, it is an empirical matter. Whether it is a good approximation depends on the world. We could change lots of different parameters by ϵ but for some parameters this will lead to a big – i.e. greater than ϵ— difference in the description of the system, but for others the description will still be similar, and accurate.

I have claimed that abstraction is involved in vertical reduction, and approximation is involved in horizontal reduction. But is this a strong dichotomy? A description of pendulum that does not mention the colour of the pendulum could be characterising as abstracting away from the colour of the pendulum, or falsely representing it as having no colour. Likewise one might think of an 'old' description of a system given by T_t as only caring about a certain level of precision. As such, it is a less fine-grained description of the system. We could reformulate the old theory as successful within a domain and to a certain degree of accuracy. This suggests that T_t has a different subject matter, a different level of precision.

If abstraction involves taking equivalence classes of lower-level states — i.e. saying

[15] I will not distinguish between approximation and idealisation here, but see Norton (2012) for a discussion of the distinction.

that certain details that differentiate the elements of this class don't matter for the higher-level theory — perhaps a similar story can be told for approximation. In the case of approximation, the equivalence classes are error margins: these results are the same within a certain error margin. There suggests that approximation is just abstraction away from certain errors, and so there does not seem to be a strong theoretical difference between the two. Are we abstracting to a different subject matter for a different purposes? Or are we just abstracting to a different level of accuracy?

One clear example where the distinction between approximation and abstraction is blurred is given (for different philosophical purposes) by Dennett (1991). In his example of a 'noisy' barcode shown in figure 1.7, the exact distribution could be replicated via a bit map. Alternatively, the pattern can be described as a barcode pattern with 25% noise. At a very coarse-grained level of description, it simply has a barcode pattern.

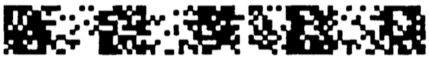

Figure 1.7: A barcode pattern with 25% noise (Dennett, 1991, p.31)

These different descriptions can be glossed in two different ways. (i) The simple 'barcode' description could be described as an approximate description: including the details of each pixel makes the description more accurate. (ii) The 'barcode' description is a higher level of description that abstracts away from the details of each pixel. The barcode can be described in a variety of ways for different purposes.

The goals of abstraction and approximation differ; and there is the following explanation for this. Abstracting to higher-levels of description — by throwing away certain details or moving a collective or new variable — can reveal new macroscopic patterns. But taking larger and larger equivalence classes of accuracy is unlikely to reveal new patterns about different subject matters, i.e. different phenomena, in way that abstraction allows us to do. Thus, whilst there are borderline cases, approximation and abstraction are different types of activity. Regardless of the noise, claiming that the pattern is a barcode is useful — for some purposes.

Whilst there is no a priori difference in kind, in particular cases we can tell the difference. The crucial difference between the theoretical devices of abstraction and approximation is whether you want to remove that device. Thus, the *goal* separates the two activities of approximation and abstraction. Is there an empirical advantage of moving to the T_b? If so, then this suggests the reduction is *horizontal*, and the theoretical device is approximation.

Of course we might get it wrong! The progress of science might reveal that we should have cared about the differences between the levels. i.e. it might be more empirical accurate to use neuroscience for the same task than folk psychology. But of course, what counts as success is context-dependent. We might want a quick way to predict something, or only need a certain level of accuracy. In which case, the details at the fringes might not matter.

1.5.4.4 Why the distinction matters: future physics won't tell us about society

Approximation and abstraction — and so horizontal and vertical reduction — can overlap, but there is a danger in failing to distinguish the two. Thus, there is a reason to make the distinction, even if it is only a difference in degree, not kind. In discussing Putnam's view of multiple realisability, Sober says "there is a lot that the physics of the present fails to tell us about societies, minds and living things. However, a completed physics would not thus be limited, or so reductionism asserts (Oppenheim and Putnam)" (Sober, 1999, p. 543).

But future physics will not be any more enlightening about societies, minds and living things than current physics. No one expects that finding (a much sought-after) theory of quantum gravity will give us insight into human behaviour. The Oppenheim and Putnam view fails to appreciate that physics has a different subject matter to psychology.

It is not that we need a better fundamental physical theory in order to describe the higher-level subject matters. To the extent, or degree of accuracy, that our higher-level theories are empirically successful, we expect them to remain so. In this sense, our higher-level theories are robust under changes of the lower-level theory. The higher-level phenomena described by T_t might be insensitive to the lower-level details. Indeed, in the case study of thermodynamics and statistical mechanics, we will see that the higher-level regularities thermodynamics describes are insensitive to the nature of matter. Likewise, Newtonian mechanics was incredibly empirically successful *in certain domains*; the advent of QM doesn't change this. Of course, there were areas where CM was not successful – small scale and very large scale phenomena. But it is only to the extent that the higher-level science depends on the details of these unsuccessful areas, that progress in the lower-level science matters. Only if biology or psychology depended sensitively on cosmological facts, will the nature of future physics matter. But inductively, we have no reason to expect this: psychology has not been sensitive to developments in cosmology.

1.6 A tool for reduction: Functionalism

Functionalism about X is the view that 'to be X is just to play the X-role'. For example, 'being locked' is classic example of a functional property: it can be realised by various mechanics — D-locks, padlocks, combination locks etc. Functionalism has become a more popular position in the philosophy of physics, so in Section 1.6.1, I first catalogue some of these functionalist views, and then I suggest this rising popularity is because functionalism is helpful for considering inter-theoretic relationships. In Section 1.6.2, I then discuss the functionalism of Lewis (1972) and Kim (1998, 1999) in the philosophy of mind, but I ultimately reject their focus on causation and Kim's eliminativism. In Section 1.6.3, I discuss what account of functionalism is appropriate for physics.

1.6.1 Functionalism in the philosophy of physics

Functionalism has become a more popular position in the philosophy of physics in recent years. As part of their wavefunction realism project, Ney and Albert (2013) are concerned to find objects in $3N$-dimensional configuration space that play the role of ordinary 3-dimensional objects such as tables and chairs. Wallace (2012b, Ch. 2) appeals to functionalist intuitions when recovering the macroscopic world from the Everrettian multiverse; for example, he claims that to be a tiger is just to be a tiger-like pattern in the wavefunction. Knox (2013) advocates functionalism about spacetime; to be spacetime is just to play the spacetime role, that is: to pick out the inertial trajectories. Functionalism about spacetime forms a springboard for considering emergent spacetime in quantum gravity (Lam, 2018).

Why consider 'functionalism' in physics? The motivation for philosophers of physics advocating functionalism about a certain concept is connected to concerns about inter-theory relations. For instance, spacetime functionalism is a position that allows us to compare spacetimes across different physical theories. Likewise, Albert and Wallace are hoping to recover the classical world that is described by classical physics, from the quantum. The key message of this Section will be that functionalism helps with reduction-as-construction: provided we can construct a quantity/equation/piece of theoretical machinery at the lower level T_b that has the same relevant behaviour, i.e. *that plays the same role* as the quantity/equation/piece of theoretical machinery at the higher level T_t, we can achieve a reduction (in this sense).

This view makes reductions easier to have. If the higher-level concepts are functional role concepts, then the realiser just has to play the same role, i.e. have the same *behaviour*. Consequently, certain differences between the quantities of T_t and T_b that one might worry block reduction — might not matter. For example, Sklar raises the following concern: The "temperature equals mean molecular kinetic energy' bridge law *identifies* a fundamentally non-statistical quantity with a fundamentally *statistical quantity*. How is this supposed to work?" (Sklar, 1993, p.161) as quoted by Batterman (2010).

Sklar's worry is that mean kinetic energy and temperature have different features: the former is statistical and latter not, and thus this blocks the reduction. But if the non-statistical nature of temperature is not part of its functional role, then the same behaviour can be captured by a statistical property: provided they have the same relevant behaviour.

This strategy originates with Lewis' plan for psychophysical identification (Lewis, 1972). This raises the question: should the lower-level property or quantity (henceforth: X_b) be identified with the higher-level property (henceforth: X_t)? That is, does the reduction establish that temperature *just is* mean kinetic energy?[16] In the next Section I outline Lewis' plan for psychophysical identification, and then Kim's functional reduction. I discuss, and reject, Kim's motivation for considering causal powers: 'ontological simplification'. I then argue that discussing causal powers is a non-starter; following Rueger I argue that for both vertical and horizontal reduction, generally X_b and X_t will have different causal profiles. I then, in Section 1.6.3, discuss 'functionalism fit for physics'.

1.6.2 Functionalism in the philosophy of mind

Lewis's plan for psychophysical identification goes as follows.[17] The higher-level concept, such as pain, is a functional role property: for a mental state to be pain is just for it to play the pain role within the whole web of folk psychology. For example, it is caused by tissue damage and typically leads to avoidance behaviour. If a state is found by physiology (putatively 'C-fiber firing') that plays the same causal role, then the occupants of the pain role would be identical to the physiological quantity 'C-fiber firing'.

Using functionalism in this way to secure reduction, or in Lewis' case, an identification, is appealing because it leaves room for the two levels T_b and T_t to differ in certain ways. For instance, one might think that if T_t is folk psychology and T_b is physiology, these levels differ in the concepts they invoke. Furthermore, our epistemic access to the concept of pain is very different from our epistemic access to the concept of C-fiber firing. The former we access through everyday, first-person phenomenal experience, whereas the latter we learn about through physiology. As such, the higher level has a certain amount of independence, or autonomy. But nonetheless, they both pick out the same state in the world. Two states can be identified, despite certain differences. (Whether this entails identifying properties requires tackling philosophically controver-

[16] One thing to flag is that in Chapter 2 I will reject this example.

[17] Lewis' functionalist ambitions are global, but here I just focus on the philosophy of mind case.

sial issues in the metaphysics of properties, such as whether properties are 'abundant' or 'sparse' (cf. Oliver (1996) for a review of the issues) — which I leave aside in this thesis.) But nonetheless Kim claims that the aim of such theoretical identifications is to reduce the number of properties one is committed to.

According to Kim (1998, 1999)'s account of functional reduction: "The model, briefly is this: (1) functionalize a higher-level property M in terms of a causal role, (2) find a "mechanism" P (the realizer of M), i.e., a property that is based on lower-level properties and can fill the causal role, and (3) find a theory at the lower level that explains how P is able to do the job of filling the causal role constitutive of M. If these steps succeed, M can be identified with P, or better: the causal powers associated with *M* can be identified with the causal powers of *P* " (Rueger, 2006, p. 336).

This model has two features that I reject. (1) The first feature of Kim's account that I reject is his view, like Lewis, that the roles are *causal* roles. Rueger (2006) argues that this strategy will not work, as the extra details at the lower level will mean that the causal roles of M and P differ. Rueger says "In more metaphysical terms: there are always causal contributions from *P* which are not needed or which are superfluous for doing the job *M* was supposed to do" (Rueger, 2006, p. 340). Moreover, since it is unclear what causation in physics is[18], functional or nomological roles are more appropriate.

(2) The second part of Kim's account I reject is his aim: to effect a 'genuine ontological simplification' by eliminating the high-level entity after identifying it with its realiser. But in Section 1.7, I will reject elimination as an aim of reduction. I now discuss how his eliminative aim motivates his use of causal roles.

The reason that Lewis and Kim focus on causal roles is because one might think following Alexander's dictum 'to be is to have causal powers'. If two quantities have the same causal powers, then perhaps they are identical. E.g. if Hesperus and Phosphorus have all the same causal powers, then this suggests that Hesperus *is* Phosphorus. In Lewis' case if pain and C-fiber firing have the same causal powers, then we can make the psycho-physical identification: pain just *is* C-fiber firing. Whereas previously we thought that there were two properties 'pain' and 'C-fiber firing', there is in fact only one. The terms 'being in pain' and 'C-fibers firing' pick out the same property as their referent.

But in the case of scientific theories, it is frequently much less clear what the referents of our scientific terms are (and indeed whether they do in fact refer in the way that the scientific realist hopes they do). Many of our scientific theories do not describe macroscopic objects like Venus or Jack the Ripper, but rather they describe complex phenomena like Bose-Einstein condensates, evaporation or quantum tunnelling. Spelling

[18] I am sympathetic to the interventionist account of causation and to the idea that causation may be a folk concept with no natural place within physics. It certainly seems hard to identify causes and effects in the pattern of events described by a time-symmetric differential equation, such as the Schrödinger equation.

out the ontological commitments associated to one's scientific models of a system is a substantive and philosophically controversial project.

Indeed this was made especially clear in List's different systems of levels. We might understand the relationships between different levels of description (or dynamics), but there is still a further project to understand how that system of levels relates to a system of ontological levels.

To look ahead to Chapter 2. Suppose we assume that the project of finding statistical mechanical realisers of the roles of various thermodynamic quantities is successful. Should we then *identify* X_{SM} with X_{TD}? Functionalism — or the functionalism I want to advocate — deflates this question. Realisation is not distinct from identification, as follows. Role-playing, rather than reference, is centre stage in the functionalist view I want to advocate. All that it is to be X is to play the X role. If David Tennant is currently playing the Hamlet role, then within this production he *is* Hamlet. In a different production, Benedict Cumberbatch plays Hamlet. Both realise the role, but it does not make sense to ask which actor we should identify Hamlet as.

This analogy with acting also illuminates our earlier discussion of multiple realisability. If to be Hamlet is just to play the Hamlet role, we then expect many different actors (i.e. 'realisers' of this role). Hence, if a theory has a functional role concepts, then it is to be expected that its concepts will be multiply realised by the lower level. We might anticipate that in different systems, different physical quantities as characterised at a 'lower level' might play a given role.

In the same way that David Tennant 'is' Hamlet: if lower-level quantity X_{SM} plays the higher-level quantity X_{TD} role, then within SM, X_{SM} 'is' X_{TD}. In this way, realisation is the most one can ask about identification. Yet, had we said that Hamlet *really is* David Tennant, like we say Hesperus *really is* Phosphorus, then we could have decreased the number of entities we are committed to — thus, achieving Kim's eliminative aim. Consequently, the role-playing view of identity is not compatible with Kim's goal of ontological simplification. But, in Section 1.7, I argue that, contra Kim, such ontological simplification or eliminativism is not the aim of reduction.

In the next Section I consider what account of functionalism is appropriate for the case studies in physics I want to consider.

1.6.3 Functionalism fit for physics

In the philosophy of mind, the functional role is understood as a *causal* role. The pain role was spelt out in terms of what typically causes it (tissue damage) and what it typically causes (avoidance behaviour). Yet it is unclear that the functionalist account from the philosophy of mind can just be imported wholesale into the philosophy of physics for two reasons. Firstly, it is opaque, at best, what causation is in physics. Secondly, much of the functionalism literature deals with issues specific to the philosophy of mind debate, especially qualia, which have no obvious analogue in physics. Given the rising popularity of 'functionalism' in philosophy of physics, it would be helpful to find an account of functionalism fit for physics.

A natural suggestion would be to find a criterion to separate the functional from the non-functional. Perhaps the obvious suggestion for such a criterion is functional definition: the functional concepts are those which are functionally defined. Indeed, a functional definition of pain is 'the property of an animal having some physical property that typically leads to aversive behaviour and is often caused by tissue damage'. The key features of a functional definition is that it is second-order and involves a pattern or web of relations. This might suggest that the property picked out by the functional definition is *extrinsic*: it depends on the existence of other properties, and of other objects, e.g. the tissues getting damaged, the limbs doing the aversive movement.[19]

However, this won't give us a way to sort the functional from the nonfunctional. How a property is defined — i.e. whether it is given a functional definition or not — does not seem to provide a criterion to help us separate functional role properties from the rest. This is because we can define *any* theoretical term in terms of the web of relations it enters into (Lewis, 1970). Given a theory, we can formulate the Ramsey sentence of that theory, where the term to be defined is the 'T-term'. The Ramsey sentence is the 'theoretical postulate' (i.e. the entire theory as a sentence) with the T-term replaced by a variable and existentially quantified.[20] In this sense, our term to be defined is implicitly defined by the whole body of the theory. Lewis illustrates this with Cluedo: A detective tells a story about a murder involving X, Y, Z, where 'X', 'Y' and 'Z' are not explained. In Cluedo the body of evidence is eventually enough to work out that X, Y, Z are Mrs. White, the billiard room and the lead piping. In this way, all scientific terms — and the properties they denote — seem able to be functionally defined by the theory they enter into.

In response, one might suggest that the difference between functional role properties and non-functional role properties is that the former can *only* be functionally defined. All properties *can* be functional defined, but only some *must* be. But proving that a non-functional definition of a given quantity does not exist seems difficult: moreover, there are reasons to think the distinction between functional and non-functional definitions

[19] However, I will not go into the intrinsic/extrinsic property distinction here because we can give a extrinsic specification of an intrinsic property. For example, 'the height of the person who is the sister of someone who is a final year medical student and cycles for Cardiff university track team and plays the oboe' picks out my height: and that is nonetheless is an intrinsic property of me, despite the relational/extrinsic specification.

[20] Really what we want is the open formula (i.e. the Ramsey sentence before it is existentially quantified), because we want an (implicit) definition which serves to pick out the functional role property in the world. The extension of the Ramsey sentence is just true or false, and so it is not quite the object we want.

will not have much bite within in physics.

This is because whether a given quantity appears functionally defined is language-dependent, or theory-formulation dependent (cf. Wilson (1985)). For example, Wilson claims that in the force-is-primitive Newtonian account, the potential energy U is functionally defined. But in Lagrangian and Hamiltonian mechanics, the potential energy U is a primitive notion, but the gravitational force "has a 'functionally defined' guise" (Wilson, 1985, p. 8). That is, it depends on which terms are primitive in the language in which it is defined. Furthermore, due to cardinality concerns Wilson claims that physics cannot have a unique language. Thus, whether a given quantity is functionally defined or not may just be a feature of the language chosen, rather than a feature of the world.

To sum up: there is not an obvious criterion to separate the functional concepts from the non-functional concepts. Indeed, I believe that Lewis' approach to theoretical terms shows that all concepts in science can be considered to be functional concepts. This formal point has an informal counterpart: "Functionalism is the idea enshrined in the old proverb: handsome is as handsome does. Matter matters only because of what matter can do. Functionalism in this broadest sense is so ubiquitous in science that it is tantamount to a reigning presumption of all science" (Dennett, 2001, p.233) (as quoted in Lam (2018)).

Of course, one might then worry that this means that 'being a functionalist about X' is a pretty empty thesis: all concepts can be considered to be functional, and claiming this about X is fairly unremarkable. Furthermore, one might expect to have lots of reductions. However, I claim that the substantive part of 'being a functionalist about X' is spelling out X's functional role. In the case of spacetime functionalism: claiming that the functional role of spacetime is to define the inertial trajectories is a a substantive position.

One might ask: is there a general prescription for spelling out functional roles? Taking the Ramsey sentence looks like an easy option. But this will not be helpful for our purposes. As we saw above, Lewis' framework uses the entire web of a given theory. But we will want a smaller part of the web, for the following reason. Functionalism in philosophy of physics is used to consider inter-theoretic relations. If the role of X is defined by the entire theory, then we will not be able to compare the X candidates across different theories. If the whole web is used then instead of having a relatively constrained role such as 'defining inertial trajectories' the entire theory will used: and so we cannot compare. But not only would using the entire web make functionalism impotent for philosophy of physics, much of the web is not important or relevant for a certain concept. (Of course the entire theoretical framework might be implicated in some way, but we want to isolate the 'essential role' of X).

Thus, substantive work in advocating functionalism in philosophy of physics is spelling out the functional, or nomological, roles. But, of course, in particular case studies, cashing out which differences matter and which don't will be very controversial. For Lewis' pain case, causal differences matter, but conceptual differences need not. In the case of thermal physics, the Gibbs vs. Boltzmann debate can be seen as a dispute over which behaviours or functional roles of thermodynamics need to be instantiated by the underlying realisers. Thus, thermodynamics defines entropy to be a strictly non-decreasing quantity: a feature captured by the Gibbs coarse-grained but not the Boltzmann entropy. However the Boltzmannian might argue that 'strictly non-decreasing' is not a crucial role of the thermodynamic entropy, cf. Callender (2001). I will focus on functional roles, but note that the terminology 'nomological' roles is also appropriate.

1.6.4 Conclusion

Functionalism can help secure reductions — but the hard work remains of spelling out the 'essential role' of the quantities in T_t. Advocating functionalism about the higher-level theory allows one to claim this theory T_t is reduced to T_b, provided that a realiser in T_b can be found (i.e. 'constructed') that plays the same role. But nonetheless the two levels can differ in certain ways. Earlier, we saw that the concepts of folk psychology differ from the concepts of physiology in our epistemic access to them. We can learn about folk psychology independently of physiology. The higher-level theory may be partially (though not wholly, due to our assumption of supervenience) independent or autonomous of the lower level. Here functionalism is illuminating: if a higher-level theory has functional role concepts, then its no surprise that the higher-level theory is *autonomous* of the lower-level theory. Mainwood (2006) gives the example of dynamical systems theory: the system is treated like a black box and so it is unsurprising that the results of dynamical systems theory are independent of, e.g. particle physics.

1.7 Metaphysics and reduction

As we saw earlier, for Kim, the aim of the reductive project is to decrease the number of properties one is committed to. But I believe that what exists is orthogonal to what is reduced, since I believe that what exists is what is useful. Newtonian mechanics provides a useful description of certain systems. Another way of putting this is that there are real patterns that Newtonian mechanics describes.[21] Consequently, I think that the entities described by the reduced T_t should only be eliminated or replaced if

[21] Some think that this 'real patterns' account of ontology is ambiguous between realism and instrumentalism, but I think that Dennett (1991)'s argument (and the Ladyman and Ross (2007) view that patterns are real if they are projectible) can be seen as a version of the no miracles argument.

they are no longer useful.

Indeed, the considerations of the previous Sections suggest that often T_t remains useful. For example, in the case of the vertical reduction, if T_t and T_b have different concerns or subject matters, then they will be useful for different purposes. But even if T_b offers an improvement over T_t (as is the case in horizontal reduction), T_t might still be useful. As we saw at the end of Section 1.5.4.1, T_t might provide a better explanation, or be more tractable than T_b for certain purposes or systems. Thus, even if T_t is reduced to T_b, T_t might still be useful and so I believe should not be eliminated. Thus, contra Kim, the slogan aim of reduction should be: vindicate! Don't eliminate!

Why think that T_t has been vindicated? Even in the case of 'improvement' reduction, T_t will have been shown — for certain degrees of accuracy, over certain timescales, and for certain systems — to give the same answers. Thus, despite no longer being the best theory for a certain subject matter or phenomena, if it has been reduced, then we have an explanation of why it was so successful. Indeed, were the theory T_t not successful, then we wouldn't care about recovering it from its successor.

I think this anti-eliminativist commitment lies behind those who advocate 'anti-reductionism' in physics. The higher-level theory T_t is useful — and furthermore in some cases knowledge of T_t might be indispensable for the securing the reduction (cf. Rueger (2006)). For instance, knowledge of the macroscopic variables might be required. At the very least, we will need to know *which* macroscopic pattern we are aiming to capture (as discussed earlier). If the higher-level information is indispensable, then it cannot be eliminated. But since I say elimination is not part of the aim of reduction, the indispensability of the higher-level information poses no block to reduction.

Of course, one might claim that eliminability is a condition on reduction — i.e. even if there are good reasons to not eliminate after reduction, one could hold that a successful reduction gives one the *ability* to eliminate the higher-level terms.[22] But I want to resist this view, because eliminability is too strong a requirement on reduction. Carrying out a reduction of T_t to T_b is a cognitive achievement, even if you need higher-level information to guide you.

Indeed, in practice we will often need the higher-level theory or concepts to guide you. To see the plausibility of this, recall just how many possible σ maps from the lower to higher level there are: that is, there are so many different equivalence classes of lower-level states. How should we know which is the right one? If you are totally ignorant of T_t, how will you find the right regularities? After all, there are so many different possible variables one can use to describe the world, as we saw in the discussion of the descriptive richness of physics earlier. Which variables we pick depends on our epistemic and cognitive limitations — i.e. on which variables we can latch onto and manipulate. Temperature is one such variable, but the centre of mass of Trump's hand,

[22] My thanks to Neil Dewar for this point.

my cat and a football is not. In horizontal reduction, T_b might be a better theory than T_t because of improvements in experimental technique and control. Thus, the variables that we can latch onto and manipulate have changed. Constructing the older T_t from T_b in complete ignorance of T_t and its variables seems like an impossible talk.

Having argued that reduction vindicates the higher-level theory, I will now discuss how reduction is nonetheless compatible with the higher-level entities being weakly emergent: indeed, I go further and suggest that in cases of vertical reduction, emergence will be endemic. I then connect this to how the higher-level entities can have their own spirit, or in other words: be autonomous.

1.7.1 Emergent special sciences

The account I have outlined in this Chapter is an account of reduction-in-practice (and, in the next Section, I will discuss the distinction between 'in principle' and 'in practice'). To declare that T_t reduces to T_b, the equations or quantities (in this latter case, functionalism may help) of T_t, must be constructed from T_b. I say construct*ed* not constructible. The construction must be demonstrated. In this sense, it is a relatively stringent requirement. But nonetheless it is compatible with (weak) emergence — understood as 'novel and robust behaviour with respect to a given comparison class' (Butterfield, 2011a,b).

If the type of reduction is Section 1.5.4.2's vertical reduction, then emergence is going to be prevalent. In fact, we might want to go as far as to say that all higher-level theories that have been reduced are, in this sense, emergent. I first suggest why they will fulfil the *novelty* criteria. I then discuss in what sense they fulfil the *robustness* criterion.

Novel: if T_t describes a genuine higher-level of description (as opposed just to being a gerrymandered level, or just a different level of 'coarse-ness of description'), then it will describe substantively different behaviour. Of course, this raises the question: what counts as a genuine level of description?

Earlier we saw that there are many different possible levels that could be defined, but some of these we might consider 'gerrymandered'. Nonetheless, it is unlikely to be fruitful for scientists to investigate these gerrymandered levels; thus, there is a sociological reason to think that the special sciences will investigated 'genuine levels'. One way of seeing that a level is genuine higher level is if it has *autonomous* dynamics, since this means that it doesn't make reference to the lower-level of description. This autonomy of the higher level is sometimes a brute fact (i.e. we just discovered the higher-level regularities in the world, like how long to brew tea) or sometimes these higher-level rules are constructed from the lower level (as we will see in Chapter 3). But in both cases, a genuine higher level will have autonomous dynamics — that is, they do not require the lower-level details. These means that certain lower-level details won't matter, and so the higher-level variables are *robust* with respect to certain changes.[23] Indeed, in List's framework, the supervenience map defines the higher-level states as equivalence classes of lower-level states — thus, by definition, it won't matter to the higher level which element of the equivalence class represents the system's state. Of course, there are *degrees* of robustness: some lower-level changes will matter.

This independence from lower-level details is one sense which the higher-level description may be said to 'have its own spirit'.[24] One example of higher-level independence: much (indeed most) of our macroscopic theories are independent of the fact that the world is quantum, not classical. But in a way this is unsurprising — insofar as these theories are empirically successful, they will remain so. But were the world classical and not quantum, matter would not be stable, and so the higher-level patterns might not emerge. As such, the independence is limited. Working out which changes don't matter — that is, which changes the higher-level theory is robust under — is part of the reductive project (cf. Section 1.4.2).

But robustness is not sufficient for emergence: there can be examples of 'good variable' choice — we can find some variable that has an autonomous description and so is robust with respect to certain lower-level changes — but it may not be suitably novel (moving to a centre of mass description might be an example).

To conclude: generally, if a higher-level description is novel and robust compared to the lower-level description, we can say that it is weakly emergent. I submit that special sciences that have been reduced will frequently earn the name 'emergent'. They are *novel* descriptions — if they were not, they would not be a genuine different level or science but would be subsumed under the science of the lower level. If a description given by the special science can be constructed from the lower-level description, then in this case we can see under changes (and so in what sense), this higher-level description is *robust*. Of course, how robust they higher-level description must be to qualify as emergent might differ across different authors. But in principle: the special sciences can be emergent, even when reduced.

1.8 Reduction in practice vs in principle

The account I have outlined is comparatively liberal; there are no constraints about explanation, for example. Additionally, the conclusion of Section 1.6 was that it seems that any theory can be 'functionalised'. Furthermore, any mathematical tools are

[23] To connect, the mathematical criterion of 'autonomy' to the issue of laws of nature: if the higher-level dynamics are autonomous, they will be time-independent. And this suggests that they may be thought of as bona fide laws of nature: regularities that are not time-dependent. Of course, this mathematical criterion for laws of nature may not be appropriate for less mathematised sciences such as biology.

[24] I will use 'independence' and 'autonomy' interchangeably to capture this idea.

permitted in trying to construct for example, the dynamical equations of one theory from another. But by allowing such liberty in my account, one might expect to find many examples of reduction; generally, the more stringent one's account of reduction, the fewer examples one anticipates finding. And yet, we do not have a plethora of examples of reduction-as-construction. Why is this?

One strategy to explain the paucity of examples, despite the liberal account of reduction: distinguish between reduction-in-practice and reduction-in-principle. Reduction-in-principle is concerned with which relations can possibly obtain between T_b and T_t. Demonstrating that these relations do in fact obtain in particular cases is reduction-in-practice. To borrow David Lewis's example, in principle I can speak Finnish. But "Facts about the anatomy and operation of the ape's larynx and nervous system are not compossible with his speaking Finnish" (Lewis, 1987, p. 77). But, in practice, I can't speak a word of Finnish.

Accounts of reduction differ over whether they have reduction-in-practice or reduction-in-practice in mind. For example, advocates of the 'limits account' such as Batterman (2001) are generalising from particular case studies of reduction, i.e. from reduction in practice — thus they are aiming to generate a principled account from particular cases. In contrast, Nagelian-type accounts that require logical deduction as a requirement on reduction, are starting from a general account of reduction, of reduction in principle — and then aiming to demonstrate that this relation holds in particular case studies, i.e. that such a reduction relation is found in practice.

The limits account focuses on mathematical relations between T_t and T_b, and mathematics is the lingua franca of physics. In contrast, the Nagelian account focuses on the logical relationships between T_t and T_b, the lingua franca for discussing possibility. In general, logical deduction is not part of the everyday practice of science.

Thus, accounts that focus on logic are more appropriate for considering reduction in principle, whereas accounts focusing on mathematics are well-suited to considering reduction in practice (in physics). Yet, mathematical approaches and logical approaches to reduction need not be in tension with one another. They both reveal interesting, but different, features of the relationships between our scientific theories. Clearly, given the emphasis on mathematical construction in my account of reduction, reduction-in-practice is my target. Indeed in the rest of this thesis I am going to consider two case studies, and so reduction-in-practice.

Thus, the difference between what we can do in principle and what we can do in practice explains why there is a chasm between the liberties allowed by the construction-as-reduction account and the dearth of examples.

Nonetheless, we might reasonably demand: why is there a difference between what can be done in principle and what we can do in practice? Fletcher gives a sociological explanation. He claims that we could find more examples of reduction — it is just this hasn't been a focus of physicists' attention. Were it to be so, we should find many more examples.

Yet I think it is not physicists' lack of interest: there are other reasons. For example, statistical mechanics considers systems of the size $\sim 10^{23}$ particles. (This is the number of molecules in one mole of gas). This is a very large number: greater than the number of grains of sand on Earth — which should give an idea of just how hard it is to solve 10^{23} coupled equations for the change in position of each particle in CM. Computational intractability makes reduction-in-practice hard.

Whilst these tractability issues are not a mere lack of interest, one might still worry that the presence of reduction — or thereof — is irrevocably entwined with our cognitive, computational or epistemic abilities. As a consequence it might seem that the dearth of examples of reduction is a mere consequence of our perspective on reality, rather than reality itself. A Laplacian demon would be able to secure the reduction, but we cannot.[25] That is, the failure of reduction in practice is an anthropocentric feature.

But there is a rich variety of patterns in the world; which ones are pertinent to us arguably depends on features of our epistemic standpoint on reality: on which variables we can measure and manipulate. Knowing the fundamental level might not be enough for the Laplacian demon, in addition to their impressive computational powers, the demon will need to know our epistemic standpoint (i.e. our cognitive limitations) in order to uncover these higher-level patterns and so effect a reduction.

1.9 Conclusion

In this Chapter I have outlined the account of reduction in practice that I will now use to consider the case studies of Chapters 2 and 3. According to my account of reduction-as-construction, if the equations or quantities of T_t are constructed from the equations or quantities of T_b, then T_t is reduced — and *vindicated* by — T_b. A whole host of mathematical devices might be required, such as defining new variables or taking limits. I claim that the use of approximations is distinctive of horizontal reduction: the reduced theory T_t describes the same behaviour of a system — to a certain degree of approximation and within a certain domain. Vertical reduction connects two theories T_t and T_b which are concerned with different subject matters, and so abstraction — the throwing away of lower-level details irrelevant for the higher-level phenomena — will be distinctive of vertical reduction. This type of reduction is familiar from the debate

[25] I have reservations about invoking the Laplacian demon. If all levels of description supervene on the fundamental level, and the demon has 'god-like' powers, he is just *by definition* able to secure reductions, which is not very enlightening. I think that this tactic is used to make it sounds as if lack of reduction is *mere* epistemological significance, and so shouldn't have any metaphysical consequences about the status of the higher-level. But earlier I claimed that reduction was independent of these metaphysical considerations, so I will not pursue this point further.

about the special sciences, which I suggested could be considered as weakly emergent. Generally, a reduction of T_t to T_b allows us to see how the macropatterns or regularities described by T_t emerge from the micropatterns and details described by T_b.

2 The reduction of thermodynamics to statistical mechanics

2.1 Introduction

In the wider philosophical literature, the relationship between thermodynamics (TD) and statistical mechanics (SM) is taken as the paradigm example of reduction. But within the philosophy of physics, there is scepticism. For example, Batterman (2010, p.159) says: "it is almost surely the case that thermodynamics does not reduce to statistical mechanics according to the received view of the nature of reduction in the philosophical literature." However, in this Chapter I will argue that according to my view of reduction: TD is reducible to SM.

Yet, as I will argue in Section 2.1.3, thermodynamics is not a dynamical theory in the way physical theories often are, and so the 'meshing type' account will not apply here. Instead, the goal will be to find the SM 'image' or realiser of various quantities. In particular I will take a functionalist approach: the SM quantities need only capture the key or essential role of the TD quantities. Of course, in spelling out these roles, temporal features, i.e. how these quantities change over time in certain interactions and situations, will be discussed. And so whilst issues about dynamics are not at the forefront, temporal notions are implicitly considered.

In Chapter 1, I argued that there is a sense in which all quantities could be considered functional. The hard part (and thus the substantive position in professing functionalism about a certain discourse) is spelling out the roles. Yet whilst functionalism could be applied to any theory, thermodynamics lends itself especially naturally to a functionalist perspective. This is because many of its core arguments and notions, such as the Carnot cycle, are very abstract. Thermodynamic systems are described by only a few parameters and the microscopic details are purposefully not considered. (Historically, the microscopic details were purposefully ignored because of ignorance: more particularly, the controversy surrounding the atomic hypothesis at the time.) In this way, functional commonality, while allowing physical diversity in the microstructure, is a theme in thermodynamics — which is conducive to taking functionalist approach.

Taking this functionalist approach will allow me to achieve two things:

1. As I outline in Section 2.1.1, it will allow me to reply to certain sceptics about reduction, by showing how TD can be reduced to SM.

2. In Section 2.1.2, I will explain the sense in which, despite this reduction, TD is an 'autonomous special science'.

2.1.1 Silencing scepticism

My views about functionalism and reduction-as-construction allow me to reply to one form of scepticism about the reduction of the TD to SM that originates with Gibbs (1903), and which Sklar (1993) endorses.

> "It should not surprise us that Gibbs, when he came to associate ensemble quantities with thermodynamic quantities in Ch XIV of his book, spoke of the "thermodynamic analogies" when he outlined how thermodynamic functional interrelations among quantities were reflected in structurally similar functional relations among ensemble quantities. He carefully avoided making any direct claim to have found what the thermodynamic quantities "were" at the molecular dynamic level." (Sklar, 1993, p.350) as quoted by (Batterman, 2010, p.161).

But according to reduction-as-construction, demonstrating the same functional interrelations — i.e. equations— between the quantities of the lower theory T_b (i.e. SM) and T_t (i.e. TD) is all that is required. Here it is tempting to ask: if an SM quantity is the lower-level realiser of the TD quantity, can the two be identified? (And one might think this project is the "were" that in the quote above). But this questions raises murky issues about property identity that I do not want to engage with, and that I set aside in Chapter 1. Instead I endorsed a functionalist view of identity: to be X is just to play the X-role. Throughout this chapter, any statement of identifying higher-level quantities with lower-level quantities should be read in this functionalist spirit. The claim SM quantity X is identified with the TD quantity Y should be read as X is the SM image/correlate/realiser of the TD quantity.

For which TD quantities do we need to find the SM realiser? Many expositions of thermodynamics explicate how the Zeroth, First, Second law of thermodynamics implicitly define new quantities (functions of state) temperature, energy and entropy respectively. For this project, I shall adopt one such exposition: Tong (2012). By following the details of such an exposition, we can thus articulate the nomological roles that these quantities play — and search for quantities in SM that have the same behaviour, and so play these roles. If this can be done across the board, that is, if all or the majority of the quantities in thermodynamics, can be constructed from, or is realised by, SM, then we have reduced TD to SM.

To summarise what lies ahead:

- In part II: The Zeroth Law implicitly defines a quantity that is numerically identical for two bodies in mutual thermal equilibrium: temperature. The SM quantity that plays this role is $\frac{\partial E}{\partial S}$. I will argue that the familiar identification — that T is mean kinetic energy — only holds in special situations and for particular systems: for cases where quantum effects are negligible.

- In part III: The First Law tells us that heat and work are interconvertible and thus there is a conserved quantity: energy. Finding an SM realiser is in some ways easy and in other ways hard: conservation of energy is a basic assumption of the microdynamics from which SM is constructed — and so SM secures the conservation of energy 'for free'. But spelling out the distinction between heat and work at the SM level is unobvious (and some have argued: anthropocentric), but I argue — following Maroney (2007)— that it can be done in quantum statistical mechanics (QSM).

- In part IV: Spelling out the nature of the Second Law is a controversial task. I argue that it should be properly distinguished from the Minus First law (whose underpinning is considered in Chapter 3). It follows that the key role of TD entropy is that it is constant in quasi-static adiabatic processes and increasing in non-quasi-static processes. I argue that the Gibbs entropy plays this role — and once again quantum considerations are enlightening.

- In part V: The Third Law does not implicitly define any new TD quantity — but I briefly discuss the sense in which it can be given a QSM, but not classical statistical mechanical (CSM) explanation.

2.1.2 Autonomy

Despite being reducible in this way, thermodynamics nonetheless has a certain degree of autonomy (and often this lies behind the scepticism about reduction, cf. (Sklar, 1993, p. 344)). But my view that thermodynamics describes functional role properties allows us to explain this autonomy (cf. Chapter 1's discussion of dynamical systems theory). Thermodynamics is autonomous from certain molecular details—exactly what one would expect of a theory that describes functional role properties. Furthermore, if you are noncommittal about the constituents of the system, you might go further and treat the system as a black box. And hence, provided that a lower-level system interacts with other systems in a thermodynamic way, it will be a realiser of the laws of TD despite the latter's lack of commitments about the internal nature of these systems.

But there are limits to this autonomy: thermodynamics is only autonomous of the lower-level theory, statistical mechanics, to a certain extent. TD is not so independent of the lower-level details that we are 'free to make any choice' about how the world could be at this level. Recall chapter's 1 assumption that the higher level (here: the TD level) supervenes on the lower level. As such, the TD is not wholly independent of the lower level. In general, a higher-level state or phenomena might be independent of some of the lower-level details — but not all.

On the other hand, there's a historical reconstruction of events according to which TD got along just fine in ignorance about atoms. Perhaps this is an oversimplistic gloss: Carnot's original cycle was inspired by a water wheel with caloric as the fluid — and so in light of the work of Joule and Thomson had to be altered. (But one might nonetheless still marvel at how slight this alteration was, given the radical revision about the nature of heat).

The independence, or autonomy, that TD has from the underlying constitution of matter is an epistemological independence. It did not matter that we did not know about the nature of matter (particles, fluids, fields?) which is hardly surprising given the 'black box' approach of TD. This is why it is frequently said that the ideas of TD apply to quantum systems, classical systems — and even black holes. (As such, the concepts are variably realised across these different domains).

In addition to explaining the autonomy of thermodynamics, this functional role understanding of thermodynamics also sheds light on controversies about the scope of thermodynamics. Thermodynamics is claimed to be a substrate-neutral theory (Rosenberg, 2008, p. 197), and this sometimes leads to the thought that its domain of applicability should thus be unrestricted; insofar as TD does not depend on the constitution of the system at hand, it should apply to all systems. This is the intuition behind the claim that thermodynamics is a universal theory, cf. inter alia Planck (1926), Eddington (1928), Atkins (2007).

But the functional nature of thermodynamic concepts puts pressure on this idea. Just because the theory is substrate-neutral doesn't mean that it is universal. Whilst TD is independent of the details of the constitution of the system, the system must nonetheless obey certain constraints — such as having equilibrium states — in order for thermodynamics to be applicable to these systems. Thus, its domain of applicability is restricted, not universal: as I discuss in Chapter 4. [1]

Yet when we are searching for SM realiser, the nature of matter does matter. As

[1] Black hole thermodynamics delivers on the state-space of thermodynamics. For example, the no hair theorems show that black holes can be characterised by a few parameters. But beyond that lies controversy. For instance, there is controversy over whether surface gravity fulfils the temperature role (cf. Dougherty and Callender (2016), Wallace (2017, 2018)). And others, e.g. Prunkl and Timpson (2018), argue that black holes do behave like thermodynamic objects — they can undergo Carnot cycles — and so are bona fide thermal objects. From the functionalist perspective described here, that Black Holes behave in the right way is all you can ask.

discussed above, one theme of this chapter will be that sometimes it is a lot easier to find the realiser or image of certain parts of TD in QSM, rather than CSM. It matters that the world is quantum, not classical. And thus this is a sense in which the nature of matter *matters* for TD. Indeed, this is unsurprising, since no higher-level regularities are independent of/autonomous from the quantum nature of matter: were the world not quantum, matter would not be stable. Thus, TD floats free from the micro-details — but not completely.

There is another sense in which TD doesn't float entirely free: its scope is constrained by the lower level. Above we saw that limits to the scope of TD are imposed internally by the theory. But limits to a higher-level theory's scope can be imposed by its lower-level realiser (or in more common jargon, by its reductive base). Indeed this is one way in which the lower-level theory is helpful: it can be used (as an additional source than experimental information) as a way of determining the scope of the higher-level theory. (Note that the same is true of SM and CM — the CM considerations of Chapter 3 show that there will be situations where the SM equations do not apply).

One might think that SM *corrects* TD in certain respects and so in this way SM constrains TD. And the lower level can be a source of information — and improvement. There is no denying that SM has the upper hand in many respects: it describes fluctuation phenomena and transport times and derives equations of state from first principles (as opposed to the empirical generalisations and phenomenological equations of state of TD). These are sources of information. We don't want to make the mistake of putting the older – if conceptually cleaner – theory on a pedestal. But whether this is the case is a source of controversy which depends on the conceptual priority of one over the other (cf. Chapter 1 and 4).

In chapter 1, I suggested that one level of description, such as psychology, is autonomous from another, such as physics, insofar as the two levels have a different subject matters. Indeed, we will see that throughout this chapter I have had to work hard to make TD and SM be 'answering the same question', i.e. describing the same situations and phenomena. As we will see in part I, each theory naturally describes different situations (SM describes: the spontaneous approach to equilibrium, the value of macroscopic quantities at equilibrium, TD describes: how quantities at equilibrium change under external interventions) suggests that they have — to a small extent — different subject matters and hence it is to be expected that TD is slightly autonomous from SM.

2.1.3 Prospectus

This Chapter consists of five parts:

- In part I, I outline thermodynamics and statistical mechanics.

- In part II, I consider the Zeroth Law.
- In part III, I consider the First Law.
- In part IV, I consider the Second Law.
- In part V, I briefly consider the Third Law, and then conclude.

Part I. Thermodynamics in general

In this Part, I describe the lay of the land according to thermodynamics and statistical mechanics, and set out the required preliminaries for what follows in the rest of this Chapter. In Section 2.2, I outline the state-space of thermodynamics: the space of equilibrium states parametrised by a few macrovariables. In Section 2.3, I tackle the controversial question about how we should understand curves through this state-space, and thermodynamic processes more generally. Because of the nature of equilibrium states, TD processes proceed by external interventions on the system, which in Section 2.4 I discuss in depth. Finally, in Section 2.5, I consider whether such interventions mark TD as worryingly different from other scientific theories. In particular, I consider whether TD is anthropocentric in the way that Bridgman (1943) argues. Then, in Section 2.6, I briefly outline the key concepts in statistical mechanics required for what follows.

2.2 Equilibrium state-space

The state-space of thermodynamics is the space of equilibrium states, parametrised by two or more macrovariables. I will call this state-space, Ξ. For a gas, the points of Ξ can be labelled by pressure and volume (p, V); for a film, they are labelled by surface tension and area; for a magnet, magnetic field and magnetization; and for a dielectric, electric field and polarization (e.g. Tong (2012, §4)).

Thermodynamic equilibrium states are states in which the macrovariables no longer vary in time: the system (as described by thermodynamics) will sit there indefinitely. Of course, the absolute nature of thermodynamic equilibrium is an idealisation.[2] Nevertheless, the key point is that we get away with treating a system *as if* it were in thermodynamic, i.e. absolute, equilibrium (at least: for the cases where TD is empirically successful.)

[2] Many features of the theories underpinning TD suggest that a system won't stay in equilibrium *forever*. For example, Poincaré recurrence suggests that systems will eventually return to earlier states. Furthermore, to take an example from (Wallace, 2015b, ft. 1): hydrogen and oxygen may seem to be in equilibrium with one another, but if you strike a match, we see the system change dramatically: that equilibrium, also, wasn't forever.

Equilibrium is at the heart of thermodynamics, and it is a presupposition of the theory that systems will end up in equilibrium. That systems will reach such an unique equilibrium state has been dubbed the 'minus first law' of thermodynamics (Brown and Uffink, 2001) — systems will spontaneously reach an equilibrium state, which then, by definition, will not change.

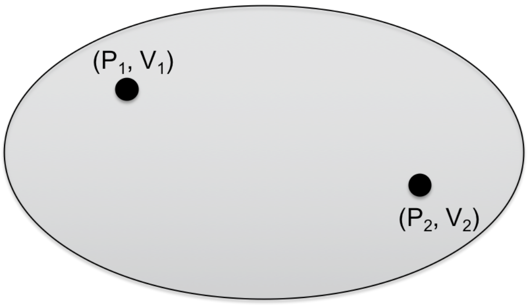

Figure 2.1: The equilibrium state-space Ξ appropriate for an ideal gas. The co-ordinates (P_1, V_1) label point x_1 and (P_2, V_2) label point x_2.

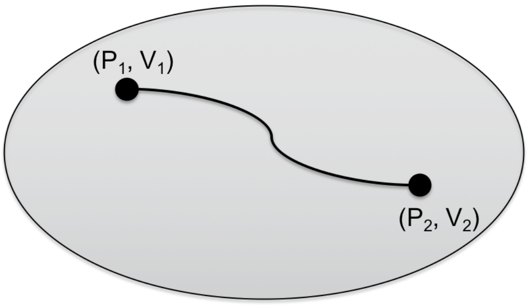

Figure 2.2: A curve through the above equilibrium state-space Ξ.

2.3 Dynamics

Having outlined the state-space of TD, we now need to consider the 'dynamics' in 'thermodynamics'. Usually, the evolution of a physical system is determined by the theory's equations of motion and its evolution can be represented by a curve through state-space parametrised by time. But this familiar situation is alien to thermodynamics. TD is not a dynamical theory. Indeed, one might think that 'thermostatics' would be a more appropriate name. There are no equations of motion and no explicit time parameter. Furthermore, its hard to see how a curve in an *equilibrium* state-space could represent any dynamical process; and hard to see which direction this process would occur.

Not only is it unclear how to interpret these curves as 'processes': Norton (2016) goes further and claims that they are paradoxical. He emphasises that the term 'equilibrium process' is oxymoronic: if equilibrium is understood to mean 'a state in which nothing changes' then by definition it contradicts a 'process' - whose meaning is that something changes; cf. also Lavis (2017) and Valente (2018). In this Section I first describe how we must conceptualise change in TD: viz. as interventions.[3] I then motivate why we must tackle the 'paradoxical' issue of equilibrium curves, before outlining my preferred position on the debate.

By the very definition of an equilibrium state, once a system reaches such a state (and so is represented by a point $x_1 : (p_1, V_1)$ such as in Figure 2.1), it will remain there indefinitely — it cannot spontaneously move to another state labelled $x_2 : (p_2, V_2)$. Thus, for any change or process to occur, there must be an intervention on the system. Its external parameters, such as its volume, must be altered: e.g. by inserting a piston.

This point is emphasised by Wallace: "[Thermodynamics] is not in the business of telling us how those equilibrium states evolve if left to themselves, except in the trivial sense that they do not evolve at all: that is what equilibrium means, after all. When the states of thermodynamical systems change, it is because we do things to them: we put them in thermal contact with other systems, we insert or remove partitions, we squeeze or stretch or shake or stir them. And the laws of thermodynamics are not dynamical laws like Newton's: they concern what we can and cannot bring about through these various interventions" (Wallace, 2014, p. 1).

But if such an intervention knocks the system out of equilibrium, then its state is no longer represented in TD state-space, Ξ. However, the Minus First Law of TD says that once the external parameter is no longer changing, the system will return to a — perhaps, new — equilibrium state.

To illustrate this, consider the following example: the Joule free expansion of a gas. The system is initially in equilibrium state x_1. The partition is removed and the gas rapidly expands in an uncontrolled manner. After some short time, the gas settles down to a new equilibrium state, x_2, with a larger volume. Only the initial and final states of this process are represented in Ξ: thermodynamics is silent on what happens away from equilibrium. Figure 2.1, but not Figure 2.2, represents the Joule expansion.

Considering a curve through the equilibrium state-space Ξ raises issues. Figure 2.2 shows an undirected, continuous curve from point x_1 to point x_2. How can such a set of points represent any process? Any intervention will knock the system out of equilibrium — indeed, this is required for anything to happen. And we can't just ignore this problem. Although many processes in TD will be like the Joule free expansion

[3] Whilst Norton has recently brought this issue to the attention of the philosophical community, others have also pointed out the problem: for example, Cooper (1967, p. 174) says these processes are 'either a contradiction in terms or limits of processes through non-equilibrium states which cannot be described in terms of equilibrium theory" as cited in Lavis (2017, p. 3).

(i.e. will not be represented by such curves), much of thermodynamics will involve examining curves through Ξ. In particular, a common strategy is to integrate the small changes in parameters such as p, V along such curves to find new thermodynamic quantities, especially ones which are path-independent. This will allow us to talk of the changes in the values of these quantities even in processes such as the Joule expansion — which involves the non-equilibrium goings-on of which TD is silent.

So let us face 'the fog of paradox' as Norton calls it. My 'defogging' strategy is to outline what I take to be the common thread to the three main recent papers on this controversy: Valente (2018); Norton (2016); Lavis (2017), who openly admit that there is not a vast difference between their resolutions.[4]

First, all agree no actual system will trace out the curve spontaneously. Hence, Tatiana Ehrenfest-Afanassjewa called these curves 'quasi-processes' to emphasise that they are unphysical, mathematical constructs (Ehrenfest-Afanassjewa, 1925, 1956). But the orthodoxy is that we can make very small interventions to external parameters, and the system will then arrive at a new, neighbouring equilibrium state and thus proceed stepwise along a curve, without ever being 'too far' from equilibrium. These small interventions are iterated, and so the system is nudged along the curve. This is Ehrenfest-Afanassjewa's concept of iterated equilibria (Valente, 2018, p. 17). The intervention takes the system away from the equilibrium, but then (due to the Minus First Law) it will return. The key idea is the deviation from equilibrium will scale with the size of the intervention. By iterating many small interventions, the system will go through a sequence of points on the curve (though not all of them). By performing more and more, smaller and smaller interventions, the system will 'stop off' at more of these points on its route from x_1 to x_2. This is embodied in Lavis' "Hypothesis of Cause and Effect: that in most situations as the manipulations of the control variables are weakened to zero, the deviation of the state from equilibrium during the ensuing process also approaches zero" (Lavis, 2017, p. 5).[5]

But we now face a problem: Norton (2016, p. 43-44) writes "Incantations of 'infinitely slow','insensible' and 'infinitesimal' have no magical powers that overturn the law of the excluded middle. Either a system is in equilibrium or it is not; it cannot be both." There are two questions: 1) how far is 'not too far'? 2) The orthodoxy is that intervening 'gently' or 'slowly enough' will ensure this closeness to equilibrium — but why should 'going slow' help?

[4] For example: "Granted, the two proposals do not seem to differ too much from each other. But it is worth to noticing that the basic intuition underlying Norton's attempt to solve the paradox was already contained in the original work by Ehrenfest-Afanassjewa on the foundations of thermodynamics" (Valente, 2018, p. 17), and "the work of this paper has similarities with that of Norton (2016)" (Lavis, 2017, p. 2).

[5] This requires that the system is thermally stable – a property that is not ubiquitous – and will be discussed in Chapter 4. An obvious counterexample, as Lavis points out, is phase transitions, where a small change in an intensive variable leads to a big change in others.

1) There is an undesirable vagueness in the claim that the system is not 'too far' from equilibrium. How far is too far? There is no satisfying answer to this question. As Valente notes, it is hard to make this precise: we can't appeal to a topology over non-equilibrium states to say that they are close enough to equilibrium, since they are not described by TD). Instead –as is so often the case with approximations– whether the system is 'close enough' is an empirical matter. Indeed, this is how Afanassjewa-Ehrenfest discussed the issue: we need an "empirically grounded concept of 'close enough to equilibrium' " (Valente, 2018, p. 17).

2) Why is it assumed that performing the interventions slowly enough will help the system stay close to equilibrium? Equilibrium requires that the macroparameters are not changing in time. The idea is that by perturbing the system *slowly* — e.g. inserting the piston slowly — the macroparameters will not be changing very quickly in time, and so the system will not be too far from equilibrium. But how should we evaluate 'fast'? Fast compared to what? There is no global, nor a priori answer, but to give a rough idea: in the case of the 'slow insertion' of the piston to intervene on the volume, the time taken to make a small change (the next 'hop' in the iterated equilibrium) should be long compared to the timescale over which the molecules bounce between the piston and the wall.

Such slow processes are called 'quasi-static'. So the picture thus far is that small interventions push the system along the curve in one direction. (A distinct intervention is required to travel in the opposite direction: e.g. removing rather than inserting a partition). The idea is that as the interventions get gentler, the deviations from equilibrium get smaller. Thus, the curve in Ξ, 'the equilibrium process', is a limit of set of non-equilibrium processes. But clearly the curve represents no process involving change. In the limit where the change to the external parameter is zero, the system remains in its original equilibrium state, and does not change. Hence this curve is a 'quasi-process'.

We have a succession of smaller manipulations leading to a succession of smaller deviations from equilibrium. But, as Lavis (2017, p. 6) says "the limit of this succession does not exist... there is no model of thermodynamics which includes the possibility of manipulation of the controls to propel the system along [the curve]. Rather [the curve] 'delimits' or is the 'common frontier' of the set of all sequences of processes" which take the system from x_1 to x_2 — via non-equilibrium states. These processes can be considered to be approximations. Thus, the term 'quasi-static' properly denotes a set of processes, whose sequence heads in the direction of the common frontier, but never meets it.[6]

The bare curve can become a directed curve — the curve can be traversed in either direction, but different interventions will of course be required for each direction. To

[6] Norton and Lavis emphasis that Duhem (1902, p. 78) has a similar approach.

go in one direction pistons must be inserted, and in the other direction they must be removed. Thus there is a set of processes whose sequence heads towards the common frontier, for x_1 to x_2, but a distinct set of processes for x_2 to x_1.

Because the curve can be transversed in either direction, there is a sense in which it is 'reversible'. But as will be discussed in Part IV on the Second Law, there are many concepts of reversibility in thermal physics. The relevant concept here is 'quasi-static': I will refer to these curves as (approximately) representing a quasi-static processes. In the special case of changing external parameters such as p or V, this is a helpful concept of reversibility. But, more generally, being quasi-static is a necessary but not sufficient as a condition on reversibility. For example, discharge of a condenser through high resistance can be forced to happen very slowly, but nonetheless it is clearly not a reversible process (Uffink, 2013, p. 277) .

2.4 Thermodynamics as control theory: Interventions

In the previous Section, I argued that the state of the system will only change when we perform certain interventions on it, e.g. inserting a partition, squeezing with a piston, placing the system in thermal contact with another, or with a heat bath... etc. For this reason, thermodynamics has been described as a control theory (Wallace, 2014). In this Section, I explains what this means.

Wallace uses the terminology 'control theory', and similar themes run throughout the foundational literature. Myrvold (2011) discusses Maxwell's means-relative view of thermodynamics, whereby certain quantities are defined relative to an agent's means. Lavis (2017) discusses a similar control theory view, but in terms of adiabatic accessibility. In the context of quantum theory, 'resource' theory views of thermodynamics are popular (Horodecki and Oppenheim, 2013b).

I believe that these foundational views bring out what is already implicit in traditional presentations of thermodynamics. Traditionally, we discuss removing a partition, or inserting a piston, or slowly varying a magnetic field. These are interventions on the system by external systems (that need not be agents in any thick sense). These interventions alter external parameters such as volume, or magnetisation — variables that would otherwise be unchanging for a system in thermal equilibrium. Hence, "all transitions between states, called *processes*, are the result of an outside intervention using a set of *control variables*" (Lavis, 2017, p. 1).

This leads us to two important points concerning 'isolation' in TD. 1) Because external systems are required in order to make interventions on the system under study, TD seems different than some other physical theories, which describe isolated systems. There must always be something else external to the system under study in order to implement these interventions. This 'something' need not be an agent, but merely some other set of degrees of freedom (DOFs), which we can call the 'controlling system'.

Clearly there is a range of interventions that could be performed. The system could be flipped upside down, shaken, stirred, or sent into outer space — but presumably not at a speed faster than light. There is a question about which interventions are possible and this will be tackled in Part IV on the Second Law. For now it suffices to note that the interventions we consider come in two types: isothermal and adiabatic. Spelling out this distinction brings us back to considering 'isolation'.

2) There is important kind of isolation particular to TD: thermal isolation. If a system is thermally isolated, it cannot exchange heat with its environment or any external system. Interventions on a thermally isolated system are called *adiabatic*.

In contrast, the system is not thermally isolated if it is in thermal contact with a heat bath. The heat bath is an object idealised to be so large that no matter how much energy (in the form of heat) flows to or from the system, the temperature of the heat bath remains the same. Such a heat bath can be used to make interventions on the system, whilst keeping the system's temperature fixed. Such interventions are called *isothermal*.

To sum up: that thermodynamics can be described as a control theory sets it apart from other physical theories which describe the space of possible states of a system and the system's *spontaneous* trajectory through that space, which is represented by a curve in that space. In the next Section, I consider whether these issues render TD 'anthropocentric'.

2.5 Interventions and anthropocentrism

The presence of manipulations or interventions in thermodynamics seems suspiciously different from the rest of our physical theories. The trajectory through phase space of a bouncing ball, or the worldline of a particle, makes no reference to which manipulations and interventions can be performed on the system. Hence, one might think these interventions are suspicious - and anthropocentric because *we* insert pistons and partitions.

Indeed, Bridgman (1943) emphasises this unusual nature of TD: he writes "[...] thermodynamics smells more of its human origin than other branches of physics — the manipulator is usually present in the argument, as in the conventional formulations of the first and second laws in terms of what a manipulator can or cannot do" (Bridgman, 1943, p. 214).

Does this 'interventions are essential' view make thermodynamics anthropocentric? I will argue: no. But, before doing so, does this question have ramifications for reduction? Myrvold (2011) suggests yes: if TD is anthropocentric then SM must be too.

Regardless, interventions need not be considered suspiciously anthropocentric. If you shift your focus, and the relevant comparison class, away from fundamental physics to other scientific theories, then the suspicion fades. In organic chemistry, we are concerned with questions such as which reactants form which products, and under what conditions particular yields are obtained. Interventions abound. Different chemicals are mixed, heated, titrated. No one is remotely worried that this, and countless other interventions that are invoked by the theories in the special sciences, is a (problematic!) anthropocentrism. Thus, I claim: in physics, the presence of another comparison class — fundamental theories with spontaneous dynamics that we take to be descriptions of the whole universe — makes 'intervention' seems suspicious. But once we see that TD is more akin to a special science like chemistry rather than a fundamental theory, suspicion fades.

2.6 Statistical mechanics construed

Statistical mechanics (SM) differs from other theories considered by philosophers of physics. Unlike quantum theory or general relativity, there are few uncontroversial axioms.[7] Instead, the situation is considerably messier: there are many different frameworks and schools in SM.

The discipline of SM is split into two parts: equilibrium and non-equilibrium statistical mechanics. Most of the foundational controversy centres around (a) non-equilibrium SM and the approach to equilibrium it describes and (b) the Boltzmannian vs. the Gibbsian approaches to SM.

In Section 2.6.1, I discuss why I am sympathetic to a Gibbsian approach to SM. Probabilities feature heavily in SM, so in Section 2.6.2 I briefly outline how probability enters both classical SM and quantum SM — and suggest that an objective understanding of probability in SM is, at the very least, plausible. Finally, in Section 2.6.3, I discuss the workhorse of SM in practice: and the connection to the two main conceptual problems of SM: (i) probabilities and (ii) time-asymmetry.

I cannot hope to do justice to the multitude of approaches: so here I outline the main contours of the debate, but much of the content of this Section will be admitting what I

[7] Uffink writes "In the foundations of quantum mechanics, one may start from the von Neumann axioms, and disregard the preceding "old" quantum theory. Statistical physics, however, has not yet developed a set of generally accepted formal axioms, and consequently we have no choice but to dwell on its history. This is not because attempts to chart the foundations of statistical physics have been absent, or scarce (e.g. Ehrenfest and Ehrenfest-Afanassjewa 1912, ter Haar 1955, Penrose 1979, Sklar 1993, Emch and Liu 2001). Rather, the picture that emerges from such studies is that statistical physics has developed into a number of different schools, each with its own programme and technical apparatus. Unlike quantum theory or relativity, this field lacks a common set of assumptions that is accepted by most of the participants; although there is, of course, overlap" (Uffink, 2006a, p. 4).

will *not* discuss.[8]

2.6.1 Gibbs vs. Boltzmann

The two main approaches to SM take the two key figures — Gibbs and Boltzmann — as their labels and inspiration.[9] The main difference between the Gibbsian and Boltzmannian approaches is that the former characterises equilibrium and entropy in terms of a probability distribution over the possible microstates of the systems, whereas the latter characterises equilibrium and entropy in terms of the microstates directly. The consensus — insofar as there is one — is schizophrenic: the Gibbsian formalism is the technical workhorse but the Boltzmannian approach is preferable when considering conceptual problems.

However, in this thesis I will take a broadly Gibbsian approach. As this is a minority view in the philosophy of physics,[10] I first give one positive reason for working with a Gibbsian perspective and then defuse one objection to this perspective, following Wallace (2013b).

The positive reason for adopting a Gibbsian framework is that this, rather than the Boltzmann framework, is the one that practicing physicists use. Thus, it is important that we make sense of the workhorse approach, rather than hiding in the conceptual niceties of an approach that does not solve so many practical problems.

Now, to defuse an objection to the Gibbsian view. The Neo-Boltzmannian approach — advocated, inter alia, by Albert (2000), Callender (2001), Price (1996), Goldstein (2001), Lebowitz (2007) — is motivated by rejecting the ignorance-based understanding of probabilities in SM, especially that espoused by Jaynes (1957). Albert rhetorically asks:

"Can anybody seriously think that it is somehow *necessary*...that the particles that make up the material world must arrange themselves in accord with *what we know*, with *what we happen to have looked into*? Can anybody seriously think that our merely being ignorant of the exact microconditions of thermodynamic systems plays some part *in bringing it about*, in *making it the case*, that (say) *milk dissolves in coffee?*" (Albert, 2000, p. 64), emphasis in original.

As a consequence, some neo-Boltzmannians, such as Goldstein (2001), aim to replace the probabilistic notions in SM with the concept of 'typicality'. But I reject this project and its motivation for two reasons: (1) I do not think that we can get away from probabilities in SM. Insofar as the intention of the 'typicality' approach of Goldstein et al. is to eliminate probabilistic concepts from SM, I believe this to be a misguided

[8] See Frigg (2010) for a state-of-the-art overview.

[9] Here I am only engaging with the contemporary debate, so following Wallace (2013b) I take an ahistorical approach. For the history of the development of statistical mechanics, see Brush (1976). For a historically informed approach to the conceptual issues, see Uffink (2006a).

[10] Other philosophical advocates of Gibbsian approach include: Wallace (2016), Maroney (2007), Prunkl (2018).

enterprise, since probability is required to capture fluctuation phenomena; cf. Wallace (2015a).

(2) We need not interpret the Gibbsian framework in a Jaynesian manner: that is, as being connected to our ignorance. Thus, whilst I agree that there is work to be done making sense of the Jaynesian approach (cf. Wallace (2013a)), this need not block taking a Gibbsian approach to SM. Wallace (2013b) helpfully summarises the situation as follows: neo-Boltzmannians object to a certain justification of the Gibbsian formalism, but that should not be confused with the formalism itself, which can be given a perfectly objective justification.

Once the Gibbsian framework is shorn of its Jaynesian justification and the Boltzmannian approach embraces probabilities, the two frameworks are not so far from one another. Indeed, Wallace (2013b, p. 2) claims that the formalism of the Boltzmannians is a special case of the Gibbsian framework. Furthermore, in an Appendix (Section 2.34), I show how the Boltzmann and Gibbs entropies are interderivable from one another.

I cannot go into further detail about the Gibbs vs. Boltzmann debate, but here is one final reassurance: whilst I will take a broadly Gibbsian approach, the success of my project does not depend on rejecting of the Boltzmannian project. Here functionalism helps: the concepts of TD can have different realisers in different theories. So if my hunch that the Boltzmannian and Gibbsian approaches are not far from one another is wrong, and the two approaches are different theories, this need not undermine my project here.

2.6.2 Introducing probability

I claimed that probabilities cannot be easily, or more importantly *usefully*, eliminated from SM. Indeed, as we will see in Section 2.6.3, probability distributions are central to the SM enterprise. Understanding and giving a comprehensive account of probability in SM is a large task. In this Section I briefly describe how probability comes into SM (in both the classical and quantum case) and I motivate why I think it is plausible that probability in SM is objective — which is all I will need for the rest of this thesis.

In classical microdynamics, the state of the system is represented by a point in phase space, Γ, which encodes the positions and momenta of the components of the system. (So for a system consisting of N point particles, the dimension of this space is $6N$).

But in statistical mechanics, "the mathematical object representing the state of the system is no longer a point in phase space, but rather a collection of points, each one being weighted by a certain number" (Balescu, 2005, p. 23). A probability density function over this phase space is a function to the real numbers in the interval [0,1] from the phase space: $\rho : \Gamma \to \mathbb{R}$. Integrating this function ρ over the phase space will give 1. Thus there is a weight assigned to each possible state of the system. This is naturally given an *epistemic* interpretation: the system is definitely in one (and only one) of the possible states, but we don't know which. But we do know that some are more likely than others and ρ quantifies this. Often this probability distribution is given a frequentist flavour by considering an infinite number of copies of the system as an ensemble.[11] The proportion of the ensemble in a given state corresponds to the weight, i.e. probability, of that state.

Whilst I said that this probability distribution is naturally given an epistemic interpretation in the classical case, this is perfectly compatible with the probability distribution being objective (as argued by Myrvold (2012)). Even though the fundamental level is deterministic, this is compatible with higher-level emergent chances, as we saw in Chapter 1 (cf. List and Pivato (2015), Butterfield (2012)). Even if one doesn't think these probabilities are deserving of the name 'chance'[12], they can be considered to be 'objectified credences' using the method of arbitrary functions, and other technical work in the foundations of probability, inter alia Skyrms (1977), Lewis (1986c), Butterfield (2011b).

Furthermore, there is reason to think that probabilistic assumptions are dynamically motivated, and so not chosen at random or in accordance with our ignorance. The entry point for probability is the microcanonical ensemble. (Henceforth, I use the terms probability distribution and ensemble interchangeably). The microcanonical ensemble assigns an equal weight to each possible state of system, under the constraint that the energy of the system is fixed: so it is confined to an energy hypersurface in the phase space, Γ. The assumption that the system is equally likely is to be in any microstate that is compatible with macroscopic constraints (such as fixed energy) is so central to SM that it is called the 'fundamental assumption' of SM (Blundell and Blundell (2009), Tong (2012)).

But this fundamental assumption need not be motivated by a principle of indifference, or sufficient reason. As I will discuss in the next Section, in order to define the microcanonical ensemble when there are continuously many possible microstates, we require the Lebesque measure which assigns a volume, and so a probability, to a region of phase space. There are many different measures that one could assign to the phase space, but the measure used in SM is dynamically motivated: the volumes assigned by this measure are invariant under the Hamiltonian flow. Thus, the way probabilities enter SM is not through our ignorance, but through dynamical considerations.

Accordingly, I think that probability can be considered to be objective in SM. However, one might worry that probability is nevertheless a new conceptual ingredient — and

[11] I think that much of the mystery-mongering about an 'imaginary infinite ensemble' is deflated by considering this frequentist reading. The real issue is not the 'ensemble', but connecting this to the outcomes of experiments via Gibbs phase averaging as discussed by, e.g. Malament and Zabell (1980).

[12] Some want to reserve the name chance for truly 'ontic', irreducible or fundamental probabilities: meaning by this probability stemming from indeterminism at the 'fundamental level'.

historically, this has been considered to be a problem for reduction. But moving from the classical case to the quantum case removes this worry. Probability is already inherent in the underlying quantum mechanics: it is not an additional ingredient. This is because the state of the system in quantum mechanics is not represented by a point in phase space but by a density matrix $\hat{\rho}$, (Baierlein, 1971, Ch. 12). As Wallace (2016) points out, a probability distribution over the fundamental microstates of QM, density matrices $\hat{\rho}$... is just another density matrix! Thus, Wallace holds that probabilities in SM stem from probabilities in QM — a position to which I am sympathetic. Note that this consolidates the problem of understanding probabilities in SM into interpreting probability in QM: thus, there is no sui generis problem of probability in QSM.

Furthermore, moving to the quantum case influences the Gibbs vs. Boltzmannian debate in favour of the Gibbsian for two reasons. (1) One of the neo-Boltzmannian objections to the Gibbsian framework is that entropy is a property of the ensemble, i.e. probability distribution, rather than a property of the individual system (Callender, 2001). But in QSM, entropy is once again a property of the system, since the density matrix is used as the fundamental description of the system. (Admittedly, there are issues, related to the measurement problem, about understanding the density matrix ρ). (2) The Boltzmann entropy is considerably less useful than the Gibbs entropy in the quantum case, since it merely corresponds to the dimensionality of the Hilbert space assigned to the system (Prunkl (2018), A. Greven (2014)).

I cannot go into the problem of probability in SM any further than this. This is regrettable, since it is one of the two main problems in SM, and is a problem worthy of a whole thesis. But, all I need for future chapters is that the probability can be considered objective.

2.6.3 The Workhorse

I now describe the workhorse of SM. Earlier I claimed that probability enters through the microcanonical ensemble. In the microcanonical ensemble, each microstate $|n\rangle$ is equally likely and is assigned the probably $\rho_{mc}(n) = \frac{1}{\Omega}$, where Ω is the total number of microstates. But this assumes that there are a finite number of possible microstates, and in CM, where q and p have continuous values, there are an infinite number of possible microstates. To accommodate this, we can talk of volumes of regions of the phase space Γ: $vol(R) = \int_R d^n q d^n p$, where $d^n q d^n p$ is known as the Lebseque measure. Thus far, we have only used the resources of multivariable calculus. The substantive move is interpreting this measure of the volume as the probability that the system is in a microstate within that volume, R.

$$\text{prob(R)} = \frac{\text{vol(R)}}{\text{vol(total relevant subvolume of }\Gamma\text{)}} \tag{2.1}$$

The probability of the total relevant subvolume of Γ that the system is constrained to is normalised to 1: the probability that the system has a microstate within this volume is 1.[13] Of course, the volume of the entire phase space may be infinite. But physical constraints on the system's state will mean that not all of the phase space will be relevant. In particular, in the microcanonical ensemble, the total energy of the system is fixed and so, the system is confined to an energy hypersurface. For the usual case of a system 'in a box' the qs are bounded and the fixed total energy means that the ps are bounded, so that the vol(hypersurface $E = k) \leq \infty$.

If we then take the system to be exchanging energy with another system (the 'heat bath'), and make natural assumptions about the possible microstates of the joint system, we can deduce the canonical ensemble for the states of the given system. (The connection between the microcanonical and canonical ensemble will be discussed in more detail in Part III on the Zeroth law).

In the case of the canonical ensemble, the energy of the system is no longer fixed. A weight $e^{-\beta E_n}$ is assigned to each microstate n that has energy E_n, where β is the Boltzmann factor $\frac{1}{k_B T}$. In order that this can be interpreted as a probability, it must be normalised. To do this, we divide by the sum of the weights of every possible microstate, $Z = \Sigma_n e^{-\beta E_n}$. Z is known as the partition function. Thus, the canonical ensemble ρ_{can} is defined as follows:

$$\rho_{can} = \frac{e^{-\beta E_n}}{Z}. \tag{2.2}$$

The canonical ensemble, like the microcanonical and grand canonical ensemble[14], has the special property that:

$$\frac{\partial \rho_{can}}{\partial t} = 0. \tag{2.3}$$

Because this probability distribution is unchanging in time, the canonical ensemble represents thermal equilibrium.[15]

The discipline of statistical mechanics splits into two parts. The first half is *equilibrium statistical mechanics*. Once we have the partition function Z we can easily find many macroscopic quantities, mainly by taking derivatives. For example, the free energy $F = -k_B T ln Z$, and the average energy $\langle E \rangle = -\frac{\partial}{\partial \beta} ln Z$. Balescu says that the partition

[13] Whilst Liouville's theorem tells us that the volume of $6N$ region is invariant under the Hamiltonian flow, a $(6N-1)$-dimensional region (i.e. a region on the energy hypersurface like we consider in the microcanonical ensemble) will only have this property if the volume is scaled by $\frac{1}{\nabla H}$. See Thompson (1972) for more mathematical details.

[14] I will not discuss the grand canonical ensemble in this thesis, but for completeness: it is the ensemble where the particle number is no longer fixed.

[15] Here I am endorsing an explicitly Gibbsian perspective, but note that Werndl, a defender of a Boltzmannian approach, claims that the canonical ensemble is so central to SM that it cannot belong to either camp (personal correspondence).

function ″contains the complete solution to the problem of equilibrium statistical mechanics. There exists, unfortunately, no such "magical" formula for non equilibrium statistical mechanics." (Balescu, 1997, p. 31). *Non-equilibrium statistical mechanics* is concerned with establishing how and under what conditions the system will reach equilibrium. If the system is initially not in the canonical (or microcanoncial, or grand canonical) ensemble, how does it end up there? This is the more controversial part of SM. Indeed, understanding the approach to equilibrium and the time-asymmetry it brings with it, is the second main problem of statistical mechanics. Chapter 3 deals with the approach to equilibrium in SM, but for now I leave the issue to one side.

2.7 The differing concerns of SM and TD

To sum up this part: equilibrium is central to TD; indeed its state-space Ξ is the space of equilibrium states. Processes in thermodynamics require outside interventions on the system: these interventions can be adiabatic or isothermal, and may proceed gently enough so that the changes to the system's state qualify as 'quasi-static'. But the concerns of SM are slightly different: equilibrium SM calculates features of the system, using a probability distribution such as the canonical ensemble. Non-equilibrium SM quantitatively describes the approach to equilibrium, which is a controversial topic. TD, on the other hand, just *assumes* that systems will reach equilibrium: and this is embodied in the Minus First Law. These 'non-equilibrium' issues will mostly be left aside until Chapter 3. But here I conclude by noting that the concerns, and so subject matters, of TD and SM are slightly different.

Part II. The Zeroth Law

Having outlined the general thesis that TD describes functional role properties, I now outline the functional role of temperature, T. Following Tong (2012, Ch. 4), I show in Section 2.8 how temperature is implicitly defined by the Zeroth law. That is, the Zeroth law is a rich enough body of information to fix that there must exist some quantity, temperature. Then, in Section 2.9, I discuss the quantity that plays this role in statistical mechanics. Next in Section 2.10, I outline Batterman's interpretation of Gibbs' objection (to the identification of thermodynamic and statistical mechanical quantities) and offer a reply. In Section 2.11, I then connect my discussion to the popular example of mean kinetic energy $\langle K\rangle$, and explain why, in general, temperature cannot be identified with mean kinetic energy.

2.8 Pure thermodynamics: the functional role of temperature

Before I show how the Zeroth law implicitly defines T, I first recapitulate the required notion of equilibrium discussed earlier. A thermodynamic description of a system involves specifying a small number of macroparameters: in the paradigmatic case of the ideal gas, the thermodynamic state of the system can be labelled by pressure and volume, (p, V). This means that all other macroparameters are functions of p and V: where the details of the function (also known as the equation of state) depends on the type of system. A system is said to be at *thermodynamic equilibrium* when its macro-parameters no longer vary in time. To see if two systems are in equilibrium with one another, we just need to put them in thermal contact with one another, and see if their states change. If they do not, then they are in equilibrium with one another. The Zeroth law of TD is:

> The Zeroth Law: If two systems A and B are each in equilibrium with a third system C, then A and B are also in equilibrium with each other.

Now we can see how this implicitly defines temperature. Assume that we have three systems A, B and C whose states are labelled by (p_A, V_A), (p_B, V_B) and (p_C, V_C) respectively. Then equilibrium between A and C requires some special relationship between the values of the different quantities defining the states of A and C, $(p_A, V_A), (p_C, V_C)$. Thus if we choose p_A, V_A, p_C then the value V_C of the volume of C is such that when A and C are put in thermal contact, nothing happens. We can write this constraint as:

If A and C are in equilibrium then:

$$F_{AC}(p_A, V_A; p_C, V_C) = 0, \tag{2.4}$$

which can be solved to give:

$$V_C = f_{AC}(p_A, V_A; p_C). \tag{2.5}$$

And likewise, if B is in equilibrium with C:

$$V_C = f_{BC}(p_B, V_B; p_C) \tag{2.6}$$

So the two expressions for V_C give:

$$f_{AC}(p_A, V_A; p_C) = f_{BC}(p_B, V_B; p_C). \tag{2.7}$$

Now we invoke the Zeroth law: if A and B are each in equilibrium with C, then A and B are in equilibrium with one another. Thus we have the constraint:

$$F_{AB}(p_A, V_A; p_B, V_B) = 0. \tag{2.8}$$

Equation 2.7 implies equation 2.8, but because equation 2.8 does not depend on p_C, this means that p_C must appear in equation 2.7 in such a way that it can just be cancelled out on each side of the equation. That is, there must be functions θ_A and θ_B and f such that

$$\theta_A(p_A, V_A).f(p_C) = \theta_B(p_B, V_B).f(p_C). \tag{2.9}$$

When this cancellation is done, we infer a relationship between the state of system A (p_A, V_A), and the state of system B (p_B, V_B), viz.

$$\theta_A(p_A, V_A) = \theta_B(p_B, V_B). \tag{2.10}$$

The value of the function $\theta(p, V)$ is the *temperature* T of the system.[16] Note that the above argument does tell us anything about the form of the function $\theta(p, V)$. (The form of T is found through the Carnot cycle). It only tells us that the property of temperature must exist. (The usual option at this point is to temporarily use the ideal gas law, $T = \frac{pV}{Nk_B}$ as a reference system to act as a thermometer, which then later gets generalised when we consider the Carnot cycle.)

Hence: temperature is a property that a system has such that if it is in equilibrium with another system (which means nothing changes when they are put in thermal contact) then they will have the same value of temperature.

A consequence of this account of temperature is that systems with very different constitutions can be in equilibrium with one another, and so have the same temperature. For instance: a photon gas, a magnet, an ideal gas and a lump of graphite can be in mutual equilibrium, and thus share the same T. This supports the earlier claim that thermodynamics lends itself to a functionalist interpretation: the above example clearly fits with the functionalist intuition that functional commonality is more important than the physical diversity.

2.9 The statistical mechanical realiser

Given that the 0th law implicitly defines the thermodynamic temperature T_{TD}, we now need to find the microphysical realiser in statistical mechanics that plays this role. The

[16] The function $T = \theta(p, V)$ is the equation of state of the system.

answer is well-known: the quantity that will play this role — that is, have the same value when two systems are in mutual equilibrium is $\frac{\partial S}{\partial E} = \frac{1}{T}$, so T is $\frac{\partial E}{\partial S}$.

That $\frac{\partial E}{\partial S}$ plays the role of T can be seen as follows.[17]

The Boltzmann entropy of a statistical mechanical system with energy E_1 is:

$$S(E_1) = -k_B ln\Omega(E_1), \tag{2.11}$$

where $\Omega(E_1)$ is the number of microstates of the system. If two systems are non-interacting the number of the available states of the joint system is $\Omega(E_1).\Omega(E_2)$.

Now we assume that the systems can interact and exchange energy.[18] The total energy remains the same so,

$$\Omega_{12}(E_{tot}) = \Sigma_{E_i}\Omega(E_i).\Omega(E_{tot} - E_i) = \Sigma_{E_i} exp[\frac{S(E_i)}{k_B} + \frac{S(E_{tot} - E_i)}{k_B}]. \tag{2.12}$$

The joint system K_{1+2} has fixed total energy so can be thought of as being in the microcanonical ensemble, where the probability of being in each state is equally likely, $p = \frac{1}{\Omega_{1+2}}$. (This is frequently called the 'fundamental assumption of statistical mechanics'). The entropy of the joint system is greater or equal to that of the sum of the two original systems, because the states of the two original systems are a subset of the total number of possible states:

$$S_{1+2}(E_{tot}) \equiv -k_B ln\Omega(E_{tot}) \geq S(E_1) + S(E_2). \tag{2.13}$$

The equilibrium argument: If the number of particles is large, we can approximate the entropy of the joint system $S_{1+2}(E_{tot})$ as follows. Entropy S is proportional to N ($S \sim N$), because the number of microstates is $\Omega \sim e^N$, and entropy S is $\propto ln\Omega$.

As seen in equation (2.12), the number of microstates of the joint system is a sum of exponentials proportional to S (and so N). But N is itself an exponentially large number ($N \sim 10^{23}$). Such a sum (over the different values of the energy of system 1, E_i) will be dominated by its maximum value $E_\star$ for the following reason. If there is a particular energy E_n such that the entropy S is twice as big for that term in the sum, so e^{2N} rather than e^N, then it will be e^N times bigger than all the other terms. As $N \sim 10^{23}$, $e^{10^{23}}$ is a *lot* bigger. Thus, this term will dominate the sum and we can approximate the value of the entropy of the joint system by this term.

How do we find this term, the energy $E_\star$ for which $S_1 + S_2$ is a maximum? It will be

[17] In what follows, I use the Boltzmann entropy for convenience, but an appendix (Section 2.34) shows how the Boltzmannian entropy can be derived from the Gibbs entropy.

[18] One subtlety: we assume that the energy levels remain the same (i.e. the interaction Hamiltonian is ≈ 0). But we should note that the existence of the interaction Hamiltonian is crucial to allow the exchange of the energy.

a maximum when
$$\frac{\partial S_1(E_\star)}{\partial E} - \frac{\partial S_2(E_{tot} - E_\star)}{\partial E} = 0. \tag{2.14}$$

As above, because this term is much bigger than the others, the entropy of the joint system is approximated well by:

$$S_{1+2}(E_{tot}) \approx S_1(E_\star) + S_2(E_{tot} - E_\star) \geq S_1(E_1) + S_2(E_2) \tag{2.15}$$

Tong (2012, p. 7) says there is no a priori reason why system 1 should have a fixed energy $E_\star$ once it is in contact with system 2, but the large number of particles involved mean its very likely to have the energy that maximises the total number of states available to the joint system.[19] Alternatively, we could appeal to Jaynes' MaxEnt principle, which states that systems will spontaneously reach a state of maximum SM entropy subject to the constraints on the system, such as fixed total energy Tolman (1938).

Why should we think that $\frac{1}{T} = \frac{\partial S}{\partial E}$ plays the role of temperature? Two systems have the same temperature, T_1=T_2, if nothing happens when they are put in thermal contact with one another. The above argument says that when two systems are put in contact energy will be transferred from one to the other so that the entropy is maximised and this happens when system 1 has $E_\star$. If two systems have equal temperatures, no energy should flow (as this counts as a macroparameter changing and so shows they are not in mutual equilibrium). And no energy will flow if the systems are already at the maximum entropy value, i.e. if system 1 already has $E_\star$ *before* thermal contact. This will be the case provided that $\frac{\partial S_1(E_1)}{\partial E} = \frac{\partial S_2(E_2)}{\partial E}$.

If temperature is defined to be $\frac{\partial S}{\partial E} = \frac{1}{T}$, then if the two systems have the same temperature, nothing will happen when they are put in thermal contact, as they are already in mutual equilibrium. Thus $\frac{\partial E}{\partial S}$ plays the same functional role as the thermodynamic temperature and so is the SM realiser of the thermodynamic temperature.

2.10 Batterman's reconstruction of Gibbs' objection

I claim that T in TD is realised by $\frac{\partial E}{\partial S}$ in SM as it plays the same role. But Gibbs was reticent about claiming to have found the statistical mechanical correlates of thermodynamic quantities, instead referring to them as analogues. Have I been too bold? In this Section I outline Batterman's reconstruction of Gibbs' worry and offer a different resolution to Batterman.

Batterman (2010) suggests that the plurality of candidates—i.e. the different ensembles of SM—might be one reason for Gibbs' reticence. Note that this is a plurality

[19] In the Boltzmannian framework, the combinatoric argument is used here.

rather than multiple realisability; it is not that the same physical quantity is realised by distinct lower-level states in distinct systems. Rather, there is more than one candidate lower-level realiser for the *same* system; as such, the contention is that the plurality is a collection of rivals.

The key idea is that the microcanonical and canonical ensembles are different probability distributions and because the SM entropy $S_{SM} = -k_B \int \rho ln\rho$ depends on the probability distribution ρ, S_{SM} differs depending on the choice of ρ. Furthermore, T appears as a parameters in the canonical ensemble, $\rho = e^{-\frac{E_i}{k_B T}}/Z$. Thus, there is seemingly a range of candidates for the statistical mechanical realisers. But are they distinct?

Batterman argues that Gibbs was unnecessarily cautious. The different ensembles are identical in the thermodynamic limit (i.e. the limit where the number of constituents goes to infinity) — and so Batterman claims that the reduction occurs in this limit. The equivalence of ensembles in this limit is meant to be evidence of a kind of 'universality': that is, that the "same thermodynamic phenomenology occurs regardless of the thermodynamic details" (Batterman, 2010, p.168). Thus, according to Batterman, we need not worry about the plurality of candidates: "The existence of the thermodynamic limit, and the demonstration of the equivalence of ensembles, provides evidence that the question of which ensemble quantity is *really* to be identified with thermodynamic entropy, say, may not even be an important question to ask" (Batterman, 2010, p.178).

Whilst Batterman claims that *which* of the plurality of ensemble quantities is the realiser of the thermodynamic quantities temperature and entropy is not an important question to ask, it is an easy question to answer. Each ensemble is appropriate to a given physical situation the system can be in. The microcanonical ensemble is appropriate when the system is isolated. The canonical ensemble is appropriate when the system is in contact with a heat bath so its energy is not fixed. Thus, in any given case there is not a plurality of candidates. Instead, only one ensemble is appropriate: it depends whether the system is in a contact with a heat bath or not.

Admittedly, physical practice is just to use whichever ensemble renders the problem tractable. That we can get away with using the two interchanging is *explained* by the equivalence in the thermodynamic limit (and the further condition that often we are dealing with systems approximately close to thermodynamic limit).

Nonetheless, one might worry that whether a system is in contact with a heat bath or not, leads to conceptually different realisers. In the next Section, I outline the link between temperature as defined via $\frac{\partial S}{\partial E}$ and as a parameter in the canonical ensemble, and thus show that the two are not conceptually distinct.

2.10.1 The connection to the Canonical Ensemble

The other place that we are familiar with seeing T in statistical mechanics is as a parameter in the canonical ensemble. However, this is not a conceptually distinct introduction of temperature, but can rather be derived from our previous definition $\frac{\partial S}{\partial E} = \frac{1}{T}$.

A system S is in contact with a heat bath R which is at temperature T. The heat bath is so much bigger than S that S can give or lose energy to the heat bath R without the temperature of R changing. We now ask: how are the energy levels of S populated in such a situation?

The number of microstates of the joint system is given by summing over the states n of R,

$$\Omega_{RS}(E_{tot}) = \Sigma_n \Omega(E_{tot} - E_n) \equiv \Sigma_n \exp[\frac{S_R(E_{tot} - E_n)}{k_B}]. \tag{2.16}$$

Because $E_{tot} \gg E_s(n)$, we can Taylor expand the entropy in the exponent and only keep the first two terms

$$\Omega_{RS}(E_{tot}) = \Sigma_n \exp[\frac{S_R(E_{tot})}{k_B} - \frac{\partial S_R}{\partial E_{tot}} \frac{E_n}{k_B}] \tag{2.17}$$

Using our earlier definition of temperature, $\frac{\partial S_R}{\partial E_{tot}} = \frac{1}{T}$, this becomes:

$$\Omega_{RS}(E_{tot}) = \exp[\frac{S_R(E_{tot})}{k_B}].\Sigma_n \exp -\frac{E_n}{k_B T} \tag{2.18}$$

Now we apply the assumption that each joint state of the joint system is equally likely. The number of joint states for which S has energy m is $\Omega_S(E_m) = \exp[\frac{S_R(E_{tot})}{k_B}]. \exp -\frac{E_m}{k_B T}$. Probability of the system S being in a state m with energy E_m is just the number of microstates with this energy divided by the total number of states (as we assume that every joint state is equally likely).

$$p(m) = \frac{\exp[\frac{S_R(E_{tot})}{k_B}]. \exp -\frac{E_m}{k_B T}}{\exp[\frac{S_R(E_{tot})}{k_B}].\Sigma_n \exp -\frac{E_n}{k_B T}} = \frac{\exp -\frac{E_m}{k_B T}}{\Sigma_n \exp -\frac{E_n}{k_B T}} \tag{2.19}$$

Here the details of the heat bath drop out and we arrive at the familiar canonical ensemble. Thus, the identification of $\frac{\partial S}{\partial E} = \frac{1}{T}$ can be used to derive the canonical ensemble we are familiar with. Thus, contra Gibbs we need not worry that the plurality of ensembles prevents us from finding an SM realiser.

2.11 Temperature is not mean kinetic energy

I've claimed that the statistical mechanical realiser of thermodynamic temperature is $\frac{\partial S}{\partial E} = \frac{1}{T}$, and that the T parameter in the canonical ensemble is derived from this. (Consequently, this is not conceptually distinct and thus does not provide a rival candidate — contra Gibbs' worries).

But the orthodoxy is that 'mean kinetic energy *is* temperature' provides the paradigmatic case of theoretical identification: and is a keystone of a Nagelian bridge law required to effect the reduction of TD to SM. But we seem to have lost sight of mean kinetic energy in the above discussion — what is the connection between $\frac{\partial S}{\partial E}$ and mean kinetic energy $\langle K\rangle$?

In fact, the connection between T and $\langle K\rangle$ substantially predates the theory of statistical mechanics with its probabilistic ensembles that Gibbs and Boltzmann developed in the decades running up to the turn of the 20th century. Bernoulli argued in 1738 that pressure of a box of gas is due to the bombardment of tiny particles on the walls as follows (Tong, 2012, p. 44). Imagine a box has side length L. A collision with the wall by a particle with velocity v_x will cause a change in its momentum $\Delta p_x = 2mv_x$ and it will hit that wall again $\Delta t = \frac{2L}{v_x}$. The force on the wall due to this atom is $F = \frac{\Delta p}{\Delta t} = \frac{mv_x^2}{L}$. Summing over all the (identical) atoms, we find:

$$F = \frac{Nm\langle v_x^2\rangle}{L}. \tag{2.20}$$

Assuming that the velocity distribution doesn't have a preferred direction, and so $\langle v_x^2\rangle = \frac{\langle v^2\rangle}{3}$ $\langle v_x^2\rangle = \frac{\langle v^2\rangle}{L}$. As the pressure is the force per area: $p = Nm\frac{\langle v^2\rangle}{3L^3} = Nm\frac{\langle v^2\rangle}{3V}$.

At this time, the ideal gas law (the infamous $pV = nk_BT$) was known as matter of experimental fact. If we equate the pressure calculated by Bernoulli with the experimentally measured pressure, we find $\frac{1}{2}m\langle v^2\rangle = \frac{3}{2}k_BT$. Thus, $\langle K\rangle = \frac{1}{2}m\langle v^2\rangle$, is proportional to temperature. This familiar case does not require the machinery of SM. Indeed, in SM we can *derive* the ideal gas law. In what follows I outline how this derivation goes — and this will then allow us to see that the connection between mean kinetic energy and temperature only holds for the special case of the ideal gas. For other systems $T \neq \langle K\rangle$ —and so temperature and $\langle K\rangle$ cannot be globally identified.

2.11.1 Derivation of the ideal gas law

We will start with the canonical ensemble, which is a density matrix:

$$\hat{\rho} = \frac{e^{-\beta\hat{H}}}{Z}, \tag{2.21}$$

where Z is the partition function. The probability that the system is in eigenstate ϕ, $p(\phi) = \langle\phi|\,\hat{\rho}\,|\phi\rangle$. If ϕ is an energy eigenstate, then equation (2.21) becomes $p(n) = e^{-\beta E_n}/Z$. Whilst we have started from the quantum description, we can find the classical partition function from this, e.g. for a single particle in phase space with Hamiltonian $H = p^2/2m + V(q)$. Further for the ideal gas, we assume there is no potential and so $H = \frac{p^2}{2m}$. The partition function for a single particle:

$$Z_1(V,T) = \frac{1}{(2\pi\hbar)^3}\int d^3q d^3p e^{\frac{-\beta\vec{p}^2}{2m}} \tag{2.22}$$

The integration of q is just the volume of the box, and the integration over $\vec{p}$ factorises into p_x, p_y, each over which,

$$\int e^{-ax^2} = \sqrt{\frac{\pi}{a}} \tag{2.23}$$

Thus, it can shown that:

$$Z_1 = V/\lambda^3 \tag{2.24}$$

where λ is called the thermal de Broglie wavelength $\lambda = (\frac{mk_BT}{2\pi\hbar^2})^{\frac{-1}{2}}$. (And the basic idea from 1924 is that $\lambda = h/p$).

For N distinguishable particles the partition function is:

$$Z = V^N/\lambda^{3N} \tag{2.25}$$

From this partition function, we can compute the free energy $F = -k_BTlnZ$, and then the pressure $p = -\frac{\partial F}{\partial V}$. Thus we get,

$$p = -\frac{\partial k_BTlnZ}{\partial V} = p = \frac{Nk_BT}{V}. \tag{2.26}$$

Thus we have derived the familiar ideal gas law.

Additionally, from the partition function we can calculate the average energy, $\langle E\rangle = -\frac{\partial}{\partial\beta}lnZ$. For the partition function in equation (2.25) above,

$$\langle E\rangle = -\frac{\partial}{\partial\beta}lnZ = \frac{3Nk_BT}{2}. \tag{2.27}$$

For the ideal gas we earlier assumed that there was no potential, so the average energy is just the average kinetic energy. Here again we see that temperature is proportional to $\langle K\rangle$.

This result can be used to show why the thermal de Broglie wavelength can be thought of as the average de Broglie wavelength of each particle. If the average energy is equated with the kinetic energy (as is done for the ideal gas),

$$\langle K \rangle = \langle E \rangle \tag{2.28}$$

For our ideal gas the right and left hand sides respectively become:

$$N.\frac{p^2}{2m} = \frac{3Nk_BT}{2}. \tag{2.29}$$

Rearranging we find: $p \sim \sqrt{mk_BT}$. In QM, the de Broglie wavelength $\lambda_{dB} = h/p$. This justifies calling the earlier quantity $\lambda = (\frac{mk_BT}{2\pi\hbar^2})^{\frac{-1}{2}}$, the thermal de Broglie wavelength.

2.11.2 "$T \neq \langle K \rangle$" in general

The above derivation of $T \propto \langle K \rangle$ from QM now allows us to see the limited domain of this relationship: T may well be no longer proportional to $\langle K \rangle$ when certain assumptions fail. One assumption above was the inter-molecular forces are negligible and so there is no potential. Clearly, this is an approximation and it only holds when the density of the gas is sufficiently low. If a gas is dense then this assumption that the interaction between molecules are negligible will no longer hold. Indeed, strong interactions are required for certain phenomena, such as phase transitions.

This assumption can be relaxed and a more realistic description given. Molecules in a gas bounce off one another. This can be modelled by including a hard core repulsion potential. The slight attraction between molecules at $r \gg r_0$ can be modelled by the Lennard-Jones potential. Incorporating these considerations leads to corrections to the ideal gas law, which are expressed as a virial expansion in terms in powers of $\frac{N}{V}$, the density. For example, we have the van der Waals equation of state: $p = \frac{Nk_BT}{V-bN} - a\frac{N^2}{V^2}$, where a contains the potential. This captures attraction at large distances and has the effect of reducing the pressure of the gas (compared to the ideal gas law). Thus, T does not depend solely on the mean kinetic energy—but also on the density, $\frac{N}{V}$.

That $T \neq \langle K \rangle$ is not solely due to ignoring the potential in the Hamiltonian in our earlier calculation. Even leaving aside inter-particle forces and the potential energy, quantum considerations show that temperature will depend on quantities other than kinetic energy. For instance, there is a correction to the pressure for a boson gas solely due to quantum statistics. First I consider the case of a gas, then solids.

Gas: Roughly speaking, the classical description (of an 'ideal' gas) will appropriate provided that the gas is neither too dense nor too cold. More precisely, (Baierlein, 1971, Ch. 9):

- $\frac{N}{V}\lambda^3 \gg 1$: the classical regime.

- $\frac{N}{V}\lambda^3 \approx 1$: the onset of quantum effects.

$\frac{N}{V}$ is the density of the gas — and is thus related to the interparticle spacing $r_0 \sim (V/N)^{\frac{1}{3}}$. Earlier we saw that the thermal de Broglie wavelength $\lambda = (\frac{mk_BT}{2\pi\hbar^2})^{\frac{-1}{2}}$. Thus, if the temperature gets low (and so the wavelength is large) or the density get high (and so r_0 is small), then the wavelength λ will be of the same order as r_0. And if λ is of the same order as r_0 then we are in the familiar case of quantum diffraction: the wave packets will spread out and so quantum effects are anticipated.

Solids: The Einstein solid model is an approximation according to which each atom is fixed in place, but vibrates in the lattice. These vibrations are treated as harmonic oscillations. We can treat a harmonic oscillator classically or quantum mechanically. Classically, the totally of a one-dimensional harmonic oscillator is

$$E_c(x,p) = \frac{p_x}{2m} + \frac{1}{2}kx^2, \tag{2.30}$$

where k is the spring constant. The contribution of the potential energy to the total energy is hardly negligible (as assumed in the derivation of $\langle K \rangle \propto T$ in the previous Section). One might try to maintain the intuitive relation between energy and temperature through the equipartition theorem, according to which the energy of each degree of freedom is $\frac{1}{2}k_BT$. According to the equipartition theorem,

$$\langle E \rangle_c = \langle \frac{p_x}{2m} \rangle + \langle \frac{1}{2}kx^2 \rangle = \frac{1}{2}k_BT + \frac{1}{2}k_BT = k_BT. \tag{2.31}$$

In this way, the potential and kinetic energy equally contribute to the total energy. T is not proportional to $\langle K \rangle$ but $\langle E_c \rangle$. In particular, details of the system at hand, such as the mass and spring constant, fall out of the picture.

But quantum mechanically, things look different. The energy eigenstates of a one-dimensional harmonic oscillator are

$$E_n = (n + \frac{1}{2})h\nu, \tag{2.32}$$

where $(n = 0,1,2,...\infty)$. ν is the characteristic frequency: $\nu = \frac{1}{2\pi}\sqrt{\frac{k}{m}}$. The expectation value of the energy is:

$$\langle E \rangle_q = \frac{1}{2}h\nu + \frac{h\nu}{(e^{h\nu} - 1)}. \tag{2.33}$$

This equation (2.33) is not a form for which the equipartition theorem is applicable (Baierlein, 1971, p. 181). The energy depends on the characteristic frequency, ν. Even if two solids, i.e. two blocks of salt, are in mutual equilibrium and so have the same thermodynamic temperature they will have different values for their kinetic energy.

Whilst the equipartition theorem is inapplicable, if $\frac{h\nu}{kT} << 1$, then the denominator may be written:

$$e^{\frac{h\nu}{kT}} - 1 = [1 + \frac{h\nu}{kT} + ...] - 1 \simeq \frac{h\nu}{kT}. \tag{2.34}$$

So in this case,

$$\langle E \rangle_q \approx \frac{1}{2}h\nu + kT \simeq kT. \tag{2.35}$$

When $h\nu \gg kT$, the energy step $h\nu$ is very small compared to the "typical thermal energy kT" (Baierlein, 1971, p.185). Intuitively this means that the quantisation of energy doesn't matter, and so we get back the classical phenomenology. But this classical phenomenology is a special —and not necessarily widespread— case: a diatomic nitrogen molecule at room temperature does not satisfy $\frac{h\nu}{kT} \gg 1$. More generally, at low temperatures this won't hold and so the equipartition theorem won't hold: here quantum effects are again important.

To sum up: whilst the infamous '$T = \langle K \rangle$' can be derived from quantum statistical mechanics, it is only true for the ideal gas. The 'ideal gas' is—as its name suggests—an idealisation. Real gases approximate the behaviour of the ideal gas under certain circumstances (the classical regime). But for cold or dense gases and for solids: $T \neq \langle K \rangle$.

2.12 Conclusion

I have shown that thermodynamic temperature is a functional role property implicitly defined by the Zeroth law of thermodynamics. $\frac{\partial S}{\partial E} = \frac{1}{T}$ is the microphysical realiser in SM that plays the same role and thus should be identified with (the reciprocal of) thermodynamic temperature—rather than mean kinetic energy, $\langle K \rangle$. As we have seen, $\langle K \rangle$ can only be identified with T for an ideal gas. For other other systems, this does not hold. As such, T might be thought of as multiply realised (but as discussed in Chapter 1, multiple realisability is no block to reduction). I considered Gibbs' reticence to identify thermodynamic T and S with SM quantities such as $\frac{\partial S}{\partial E}$ due to the plurality of candidates, and agreed with Batterman that this hesitance is unnecessary — but I offered a different solution to Batterman.

I now must consider the First, Second and Third laws of thermodynamics in order to see if the success had with temperature can be had with other thermodynamic quantities. If this is so, then I claim: thermodynamics can be reduced to statistical mechanics.

Part III. The First Law

The First law is seemingly trivial. Section 2.13 outlines the phenomenological statement of the First Law. Section 2.14 briefly outlines the historical context, which explains why this law was such an achievement — it established the existence of a thermodynamic quantity called energy, which is conserved.

In Section 2.15, I consider where this leaves us with respect to reduction. On the one hand, Sklar (1993) argues that reduction is easy: heat and work are just energy. But following Uffink (1996), I explain why the distinction between different kinds of energy transfer is central in TD. In Section 2.16, I then describe why cashing out the distinction between work and heat looks difficult: from the lower-level perspective, both are 'nought but molecules in motion'. This leads to the worry that heat and work are 'only defined in relation to the mind which perceives them' (Maxwell, 1871) — that is, the worry that anthropocentrism abounds. But Section 2.17 dispels these worries: we can cash out the distinction in QSM without anthropocentric worries.

2.13 Phenomenological Statement of the First Law

Tong (2012, §4.2) states the First Law as follows:

> The amount of work required to transform an isolated system from state 1 to state 2 is independent of how the work is performed.

No matter how the work is performed — stirring, squeezing, passing a current through it — the change in energy of an *isolated* system is equal to the work done $\Delta E = W$. Tong says that "this rather cumbersome sentence is simply telling us that there is another function of state of the system, $E(p, V)$." (Tong, 2012, p. 111).

If the system is *not* isolated, then the energy change is not just equal to the work done. For instance, if two systems at different temperature placed in thermal contact, the energy of the colder temperature will increase. This transfer of energy is called 'heat'.

Putting these components together we have: $\Delta E = Q + W$. Importantly, heat and work are not forms of energy, but of energy transfer. As such it doesn't make sense to say that the system 'contains', or 'has' a certain amount of heat or work. Consequently, whilst energy is a function of state so that an infinitesimal change in energy can be written as a total derivative:

$$dE = \frac{\partial E}{\partial p} dp + \frac{\partial E}{\partial V} dV, \tag{2.36}$$

the same is not true of heat and work. Thus, we denote 'small changes' in heat and work by inexact differentials as follows:

$$dE = đQ + đW. \tag{2.37}$$

From our modern perspective, the First Law looks trivial; it merely states the conservation of energy — a cornerstone of modern physics so central that violations seem unimaginable. To us, energy dE is the quantity most well understood in equation (2.37). It is gaining an understanding of heat and work that is the epistemic challenge.

But it was not always so. Historically, heat and work were the more well-understood, or secure, concepts in equation (2.37). In the next Section, I describe how heat was previously considered as an invisible fluid (and consequently a body could 'contain' a certain amount of heat, contra our modern understanding). It was the interconvertibility of heat and work which established the existence of a conserved quantity, energy, which the First Law states is 'fungible'.

2.14 A historic achievement

By the late eighteenth century, much had been experimentally established about the heat capacities of different substances. Through experiments by the likes of Black and Fahrenheit, heat had been shown to flow between bodies of differing temperatures (Cercignani, 1998, Ch. 2). Furthermore, the fact that heat could be transformed into work—paradigmatically in a steam engine— was well-known and indeed, was powering the industrial revolution.

But the metaphysical picture of the nature of heat was vastly different than our picture today. Lavoisier's theory was that heat was an invisible fluid — composed of elementary particles. This theory could explain many phenomena: the particles of the caloric fluid repelled one another, so the fluid would spontaneously flow from hotter to colder bodies. Furthermore, the increase in temperature of a gas that is being compressed could be explained as the increased density of caloric fluid.

However, Lavoisier's ideas hit a roadblock.[20] Thomson, later known as Count Rumford, was overseeing the Munich artillery whilst thinking about Lavoisier's caloric fluid. Boring cannons generated heat, which was explained by the caloric theory as being due to the pressure and movement of the caloric fluid being squeezed out, especially from the metal shards that fragmented off the cannon. But there was no detectable differences between the shards and ordinary brass.

Moreover, the crucial part for casting doubt on the caloric theory that Rumford emphasised is that the source of the heat seemed inexhaustible. Had the caloric theory been correct, the expectation was that the caloric would run out: "the source of the heat generated by Friction, in these experiments, appeared to be evidently inexhaustible" (Rumford, 1798, p. 99). Thus heat "cannot possibly be a material substance..." and he concludes it must be "motion".

[20] Lavoisier was beheaded during the French revolution in 1794; Rumford married his widow.

The final piece of the puzzle was established by James Joule.[21] Joule discovered that by stirring water, the temperature would increase: "the mechanical equivalent of heat" Joule (1850). Thus, work could be transformed into heat. Thus, the inter-convertibility of heat and work was established. Heat and work are thus seen to be different causes of a system's energy changing. Energy conservation, previously thought to be a mere curiosity of particular mechanical systems with certain symmetries, was elevated to a law of nature.[22]

2.15 Done and dusted?

Heat and work are both forms of energy transfer: does this mean that reduction is easily had? Certainly there seems to be no ontological problem. Sklar (1993, p. 349) suggests that in the same way that visible light studied by the theory of optics was discovered to be an electromagnetic wave, heat has been discovered to be a form of energy transfer.

Is the law of conservation of energy the microphysical image of the First Law? In a certain sense, yes: if energy is conserved, then the First Law holds. But the First Law also draws a distinction between heat and work, in a way that mere conservation of energy does not.

In purely phenomenological thermodynamics, there's already an understanding of what heat and work are — they are familiar from previous theories. Thus, from the perspective of 1750-1840, it is the conservation of energy that was the epistemic achievement or informative part of the First Law. But from the perspective of SM, the conservation of energy is fairly trivial. This is because SM is constructed from the underlying Hamiltonian mechanics by adding probabilistic concepts, as we saw earlier.[23] Energy conservation is already a feature of these microdynamics, and SM doesn't have to do anything to ensure energy conservation: it is already a baseline assumption of the microdynamics underlying SM.[24] The difficult problem from the perspective of SM, is gaining an understanding of the distinction between heat and work. In this Section I explain why this distinction — in some guise or another – is required.

Whilst the concept of heat Q is central in the original derivation of the thermodynamic entropy, S: $dS = đQ/T$, one might nonetheless hope to do away with heat in thermodynamics. Indeed, Carathéodory (1909) aimed to eliminate heat as a fundamental concept in TD — precisely because he thought it could be defined in terms of energy

[21] Whether Joule or Meyer discovered this first is contested, see Brush (1976) for more details.

[22] See e.g. Elkana (1974) and Kragh (2001) for more details.

[23] Of course, other assumptions will be required: as will be discussed extensively in Chapter 3.

[24] Note that if we view the situation not from the SM perspective but from the perspective of the underlying Hamiltonian mechanics, then the energy conservation resulting from the time-independence of the Hamiltonian is non-trivial.

(Uffink, 1996). Carathéodory's framework is far more rigorous than other formulations; and was the historical seed of the 'axiomatic' approach to thermodynamics (cf. as recently represented by Lieb and Yngvason (1999)). Whilst Carathéodory dispenses with heat, the distinction between adiabatic and non-adiabatic is crucial: Hornix (1970) highlighted that on Carathéodory's approach, the term 'adiabatic' becomes an unexplained primitive in the theory (Uffink, 1996, p. 384). Previously our understanding of 'adiabatic' was dependent on 'heat': an adiabatic processes is an thermally isolated processes, i.e. no heat can flow to the system.

Thus, the term 'heat' has been eliminated, but a related distinction 'adiabatic vs. non-adiabatic' is essential. Indeed, as will become apparent when we consider the Second Law, the distinction between adiabatic quasi-static processes and non-adiabatic quasi-static processes is crucial in TD: because the former are 'entropy-neutral' in a way the latter are not. (Again, this echoes the idea that the concept of 'thermally isolated' as the most fundamental to all of thermodynamics' (Kestin, 1979, p. 72, Vol. 1) as cited by (Brown and Uffink, 2001, p. 528)).

To see that *some* such distinction — that is, there is an asymmetry between heat and work, or thermal and mechanical variables, or adiabatic and non-adiabatic, is fundamental in TD, consider the following argument (repurposed from (Uffink, 2001, p. 370)[25]). The validity of Carathéodory's formalism is invariant under the pairwise permutation of the meanings 'heat/work', 'thermal/deformation coordinate' and 'adiabatic/without exchange of work' (Uffink, 2001, p. 370). An analogous expression to $dS = đQ/T$ is found: $đW = pdV$. The Carathéodory framework allows us to make sense of irreversibility in this 'pairwise permuted' interpretation as follows: a system with positive pressure can, without doing work, increase the volume by expansion into a vacuum, but one cannot decrease the volume without doing work. But the 'pairwise permuted' interpretation has empirical differences from our original interpretation: "a fluid with low pressure can very well do work on another fluid with high pressure by means of a lever or hydraulic mechanism" (Uffink, 2001, p. 370). But a system with a low temperature cannot transfer energy in the form of heat to a system with high temperature. Hence, if there was no important distinction, the permutation pairwise should not be significant. But it *is* significant: it leads to a concept of irreversibility that is not empirically adequate — and thus, the distinction is important.

To sum up: there seems to be no getting away from the fact that we need to distinguish different types of energy transfer in TD. Uffink concludes that "what we are left with is that, in some guise or another, the distinction between types of energy transfer, whether we call them heat vs work, or a transfer between adiabatic vs non-adiabatic walls,... is

[25] Uffink's original purpose was to show that the content of the Carathéodory's Second law is not identical to the Planck, Kelvin or Clausius formulations of the Second Law: exchanging the meanings retains the validity of Carathéodory's formalism, but leads to an expression not equivalent to the Kelvin or Clausius formulations.

essential to the structure of thermodynamics" (Uffink, 1996, p. 384). I claim that this means that the heat/work distinction is an essential part of the 'energy role' in TD — conservation of energy is not enough. Thus, the project of finding the SM realiser is not done and dusted — yet.

2.16 Molecules in motion

The distinction between heat and work is central to TD. But when we look at the molecular level, finding a reflection of this distinction is not obvious; both heat and work correspond to 'molecules in motion'. Hence, "the distinction between work and heat must be something like the distinction between the kinetic energy associated with random, chaotic motion of molecules, and the kinetic energy associated with aggregate motions of molecules. Even if we can find some way of carving these up, it's hard to see how this could be a natural distinction; after all, from the perspective of the kinetic theory, it's all just kinetic energy" (Knox, 2016, p. 56).

At this point, it might be tempting to adopt the following defeatist attitude: the concepts of heat and work belong at the higher-level of description. What seems like a natural carving of the conceptual landscape in the macroscopic realm of steam engines, gases and large magnets may fail to have an obvious correlate. The worry is that the lower-level candidates are not fit for the job; the distinction between ordered and disordered motion does not capture the heat/work distinction.

In Section 2.16.1, I first consider whether this is because the heat/work distinction is only applicable to the macroscopic realm — and I claim that it is not. I then consider the worry, originally raised by Maxwell, that the distinction is anthropocentric. Ultimately, I suggest that we can discard the disorder/order distinction as the microphysical image of the heat/work distinction, and in Section 2.16.3 I outline why SM can do better. In Section I outline what the image of the heat/work distinction is in *statistical mechanics*, rather than merely at the 'microscopic level'.

2.16.1 Macro/micro

Dissatisfaction with the distinction between disordered and ordered motion could suggest that heat and work are not concepts that apply to the microscopic realm.[26] For instance, it seems like a category mistake to ask what the colour of an electron is. Likewise, one might think that asking the temperature of a single molecule at an instant of time is also a category mistake.

[26] If the concepts of heat and work only have a very limited domain, then it seems there will be consequences downstream for the Second law, since it is defined in terms of terms of heat and work. (The scope of the Second law will be considered extensively in Part IV of this Chapter).

It does seem that we cannot state the amount of heat or work associated to a single molecule at an instance of time. But this point is hardly surprising: we stated at the outset that heat and work are forms of energy *transfer* — thus, they are defined over a period of time.[27] Temperature, heat and work — like many concepts— only apply over certain timescales. Do they also only apply over certain length scales? That is, do these concepts only apply to systems over a certain size. Ultimately, a concept's domain of applicability depends on the extent to which it is useful. Szilard imagines a one-molecule gas: if a single molecule is confined to a box and we consider its behaviour over a long enough period of time, then it seems like we can define its temperature and other macroparameters. Given we can apply the concepts of heat and work to a one-molecule gas (cf. Szilard (1929) and the ensuing literature e.g. Leff and Rex (2002)), the concerns of this Section don't seem to require a limitation to 'large N' system.

2.16.2 Anthropocentrism?

The distinction between ordered and disordered motion seems subjective: what looks ordered and tidy to one person may look messy and unruly to another. This introduction of anthropocentrism was suggested by Maxwell:

> "Available energy is energy which we can direct into any desired channel. Dissipated energy is energy we cannot lay hold of and direct at pleasure, such as the energy of the confused agitation of molecules which we call heat. Now, confusion, like the correlative term order, is not a property of material things in themselves, but only in relation to the mind which perceives them. A memorandum-book does not, provided it is neatly written, appear confused to an illiterate person, or to the owner who understands it thoroughly, but to any other person able to read it appears to be inextricably confused. Similarly the notion of dissipated energy could not occur to a being who could not turn any of the energies of nature to his own account, or to one who could trace the motion of every molecule and seize it at the right moment. It is only to a being in the intermediate stage, who can lay hold of some forms of energy while others elude his grasp, that energy appears to be passing inevitably from the available to the dissipated state" (Maxwell, 1878, p. 221); (Niven, 1965, p. 646) as quoted in Myrvold (2011)).

There is more than whiff of anthropocentrism here. Indeed, this is surely the type of comment that leads to Bridgman's view that TD "smells of its human origins more than

[27] Whilst there are some concepts which seem to be ubiquitous —energy, mass— these seem to be the exception not the rule.

other physical theories" (Bridgman, 1943, p. 214). But at which concept's feet should we lay the charge of anthropocentrism?

There is a question of *which* of the theory's concepts look anthropocentric, according to Maxwell. In the quote above, 'disordered motion' is the seemingly anthropocentric concept. Thus, it seems that the *image* of the TD heat/work distinction is anthropocentric — i.e. it is the lower level theory that contains an anthropocentrism, rather than TD itself. Once again we run into the issue of the purpose/account of reduction familiar from Chapter 1. The anthropocentrism only 'spreads' or 'scales up' to TD, if we think that SM tells us what the TD concepts *really were* along. But, as discussed in Chapter 1, I am not necessarily committed to the quantities of the higher-level theory T_t (here: TD) being *replaced* by the quantities of the lower-level theory (here: SM).

Furthermore, functionalism allows that there can be certain differences between the two theories; the quantities of the reduced theory T_t needn't 'inherit the natures' of the quantities of T_b: here, the SM and TD quantities can differ over whether are anthropocentric or not.[28]

Regardless, we can leave these concerns to one side, because we can do better than claiming that the image of the heat/work distinction is the disordered/order motion distinction in Section 2.17. The next subsection explains why we might think we can do better.

2.16.3 More than molecules in motion

At this point, one might be tempted to ignore the heat/work distinction and side with Sklar: SM has energy conservation and *that* is the image of the First Law. But Section 2.15 claimed that the heat/work distinction was an essential part of the role.

Agreed, merely looking at the 'microscopic level' at molecules jiggling around does suggest that both heat and work are "nought but molecules in motion" (Maxwell 1874) as cited in (Uffink, 1996, p. 373). But as Uffink emphasises, there are more 'lower-level' resources available than the bare description of the motion of molecules. "Statistical mechanics has more at its disposal than the concepts of mechanics and the molecular view alone, and one can reasonably expect that the distinction between heat and work can be framed in a mind-independent way with the help of concepts from probability theory" (Uffink, 1996, p. 344): i.e. by considering SM, we can do better than the distinction between disordered/ordered motion.

Uffink says that it does not suffice to identify heat with a form of energy transfer: we want to know *which* form of energy transfer. Here he notes the difference from Sklar's preferred exemplar of successful reduction, the case of optics and electromagnetism. In that case, we can not only identify visible light with electromagnetic waves, but also

[28] Furthermore, recall from Part I that TD needn't be considered anthropocentric.

specify *which* part of the electromagnetic spectrum corresponds to visible light.

To sum up: looking at the microscopic realm, it seems hard to see the distinction between heat/work — and distinction we do see looks anthropocentric. But we have more resources available in SM than merely looking at jiggling molecules. I now consider the realiser of the heat/work distinction in SM.

2.17 The image in SM: quantum heat and work

Both heat and work are forms of energy transfer. Earlier I claimed that the distinction between heat and work was an important part of the nomological role of energy. Additionally, I also claimed earlier that probability in SM would help us find a more precise image of heat and work than disordered/ordered motion.

In classical statistical mechanics, heat is $đQ = E_i dp_i$ and work is $đW = p_i dE_i$. One gloss on this: the changes to the external parameters, i.e. the work done, alter the energy associated to state i. But the heat flow changes how those energy levels are occupied (as it changes the probability of the system being in state i with energy E_i, because the system has more energy and so the probability of higher energy levels being populated is increased.)

Of course, this then raises the question: how should we understand these probabilities? If they are understood to be a measure of our ignorance, then perhaps the charge that the lower-level images of heat and work are anthropocentric returns. However, I am going to leave aside the CSM cases, and focus on QSM. This is because I find it the more perspicuous framework, since how the external parameter is changed comes into the quantum Hamiltonian, e.g. the volume of the box determines the energy eigenstates, while it is less directly obvious how changes to external parameters affect the energy of a classical gas — as there is no potential.

Here I show how the different changes to external parameters, i.e. mechanical coordinates, alter the probability distribution ρ — and thus, following Prunkl (2018) and Maroney (2007), find the quantum expression for work. Heat flow requires that the system not be thermally isolated; I outline how the interactions with an environment give rise to an expression for heat.

The work done is the energy change due to interventions on the external parameters, or mechanical coordinates, of the system, such as the volume. In contrast, the energy transfer due to heat flow changes the thermal coordinates such as the temperature. Of course, there are many ways to heat or cool a system. But Baierlein claims that what they have in common is that "the external parameters remain fixed, but there is nonetheless an interaction with the environment that leads to a transfer of heat...the energy thus transferred we call heat" (Baierlein, 1971, p. 205).

First I outline the strategy for work, and then heat. One disclaimer at the outset: as is familiar from ordinary QM we will be dealing with expectation values of observables such as energy given the quantum state of the system, ρ.[29]

Work: The mean energy of the system whose density matrix is ρ is $\langle H\rangle_\rho = TrH\rho$. Changing the external parameters, such as volume, changes the Hamiltonian governing the system. The Hamiltonian depends on the external parameters such as volume, as follows.

A gas confined to a box can be modelled by the familiar quantum 'infinite square well', for which:

$$V(x) = \begin{cases} 0 & -a \leq x \geq a \\ \infty & |x| > a \end{cases} \tag{2.38}$$

where the length of the box is $2a$ (this is the one-dimensional case but it can be easily generalised to three dimensions). This imposes the boundary condition that $\Psi(-a) = \Psi(a)$. This imposes that the allowed wavelengths are $\lambda_n = \frac{4a}{n}$, where $n = 1, 2, 3....$ The volume of the box (or in this one-dimensional case: the length of the box) enters into the allowed eigenvalues: $E_n = \frac{\hbar^2 k_n^2}{2m}$ where $k_n = \frac{n\pi}{2a}$. Thus we can write that the Hamiltonian depends on the volume $H[V]$. Of course, in other cases it could depend on different external parameters.

For now, we take the system to be isolated; in this context 'isolated' means not interacting with any other system, rather than a time-independent Hamiltonian. The Hamiltonian might be varied over time, for example, through the manipulation of external parameters. For such an isolated system, the change in the average energy is:

$$\frac{\partial\langle E\rangle_\rho}{\partial t} = \left\langle \frac{\partial H}{\partial t} \right\rangle_\rho \tag{2.39}$$

where ρ is the density matrix describing the system's state.[30]

Maroney (2007) and Prunkl (2018)'s analysis then integrates equation (2.39) (and demands a cyclic variation of the Hamiltonian: $H = H_0$ for all $t \leqslant 0$ and $t \geqslant \tau$.

$$W = -\int_0^\tau \frac{\partial\langle H\rangle_\rho}{\partial t} dt \tag{2.40}$$

The term W is earns the name 'the work done', since the system is isolated, there is no other energy flow to the system. So the work done is the change of the expected

[29] Naturally, there are interpretative issues concerning the probability inherent in this description. But to reiterate my earlier strategy: these issues about probability are general issues in QM, rather than sui generis issues for quantum *statistical* mechanics.

[30] I take this to be the canonical ensemble, but it is interesting to note that Maroney's analysis does not depend upon this assumption. For an explanation of the RHS of equation (2.39), see (Maroney, 2007, p. 15)

energy

$$W(\rho) = Tr[H(\rho(t) - \rho(0)] \tag{2.41}$$

Heat: Now, let's consider heat flow. For the energy transfer we call heat to occur, the system must be put in thermal contact with an environment, i.e. a heat bath. Thus, there must be an interaction term in the Hamiltonian to describe the interaction between the system and environment. The system is no longer isolated. Baierlein says "there must be an additional term in the quantum mechanical operator for the system, a term that couples the system to the external world and permits an exchange of energy despite the constancy of the external parameters" (Baierlein, 1971, p. 205).

We can describe the system and environment (or heat bath) initially as $\rho = \rho_s \otimes \rho_e$: this means that they are initially uncorrelated. This joint system is governed by the following Hamiltonian:

$$H(t) = H_s \otimes I_e + I_s \otimes H_e + V_{se} \tag{2.42}$$

where H_s describes the changes to the external parameters of the system (if any), likewise H_e for the environment. V_{se} describes the interaction between the system and environment.

We can deduce the following expression: [31]

$$\left\langle \frac{\partial H}{\partial t} \right\rangle_{\rho} = \left\langle \frac{\partial H_s}{\partial t} \right\rangle_{\rho_s} + \left\langle \frac{\partial H_e}{\partial t} \right\rangle_{\rho_e} + \left\langle \frac{\partial V_{se}}{\partial t} \right\rangle_{\rho} \tag{2.43}$$

which (Maroney, 2007, p. 17) says is unsurprising: "it tells us that the rate at which the mean energy of the combined systems changes equals the mean rate of work performed on each of the two subsystems plus the interaction between them".

With some re-arranging, Maroney and Prunkl define a term Q which describes the energy change due to the interaction, rather than changes to external parameters. Thus, they deduce:[32].

$$i\hbar \frac{\partial \langle H_s \rangle_{\rho_s}}{\partial t} = \left\langle \frac{\partial H_s}{\partial t} \right\rangle_{\rho_s} + Q[H_s]. \tag{2.44}$$

And likewise for the environment:

$$i\hbar \frac{\partial \langle H_e \rangle_{\rho_e}}{\partial t} = \left\langle \frac{\partial H_e}{\partial t} \right\rangle_{\rho_e} + Q[H_e] \tag{2.45}$$

[31] This seems to involve assuming the system and environment system is not interacting with a third in order to go from $\frac{\partial \langle H \rangle_\rho}{\partial t}$ to $\left\langle \frac{\partial H}{\partial t} \right\rangle_\rho$

[32] See equations 26-28 in Prunkl (2018) or equations 80-82 Maroney (2007)

$$i\hbar\frac{\partial\langle V_{se}\rangle_\rho}{\partial t} = \left\langle\frac{\partial V_{se}}{\partial t}\right\rangle_\rho - Q[H_s] - Q[H_e] \tag{2.46}$$

where $Q[H_s] = \langle[H_s, V_{SE}]\rangle_\rho$ and $Q[H_e] = \langle[H_e, V_{SE}]\rangle_\rho$. "The term $Q[H_s]$ clearly represents the mean rate at which energy is flowing into the system in addition to the work performed on it" (Maroney, 2007, p. 1). [33]

The two systems are taken to not to be interacting before $t = 0$, and to not be interacting after τ. (This is the cyclic variation of the Hamiltonian discussed earlier). From this, (Prunkl, 2018, p. 31) defines:

$$\Delta E = \int_0^\tau \frac{\partial\langle H\rangle_\rho}{\partial t}dt = \langle H(t)_{\rho(t)}\rangle - \langle H(0)\rangle_{\rho(0)} \tag{2.47}$$

$$\Delta W = \int_0^\tau \left\langle\frac{\partial H}{\partial t}\right\rangle_\rho dt \tag{2.48}$$

$$\Delta Q = \int_0^\tau Q[H_s]dt = \int_0^\tau Q[H_e]dt \tag{2.49}$$

which then gives: $\Delta E_s = \Delta W_s + \Delta Q$ and $\Delta E_e = \Delta W_e - \Delta Q$: the familiar First Law. To add to the idea that these quantum expressions do play the heat and work role, note that there is an upper limit on how much energy can be extracted on average as work from the system.[34] This is what is known as the adiabatic accessibility: "not all of the energy of a system is available for work" (Maroney, 2007, p. 22).

2.18 The First Law: Conclusion

The First Law states that a change in energy is the sum of the work done on the system and the heat flow to the system. Energy can be transferred to or from the system in forms but the total energy is conserved. Capturing the image of the conservation of energy at the lower-level looks trivial: the conservation of energy is a baseline assumption of the microdynamics from which SM is constructed.

But I argued that there is more to the energy role in TD than being conserved: the First law makes a distinction between heat and work. There is no escaping the importance of such a distinction; even in Caratheodory's framework which eliminates the term 'heat', there is an essential distinction between 'adiabatic' and 'non adiabatic'. Thus, the SM realiser is not the mere conservation of energy — we also needed to find the image of the heat/work distinction in TD.

[33] Here I have only provided a sketch of the extensive work of Maroney et al. For example, Maroney extends this analysis by considering the particular forms the interaction hamiltonian can take.

[34] There is a further condition: the Hamiltonian must be bounded from below.

One candidate is the distinction between ordered and disordered motion — which Maxwell branded anthropocentric. But I argued that we can do better: I claimed that a realiser of heat and work is found in quantum statistical mechanics.

Part IV. The Second Law

2.19 Introduction

The Second law is the most well-known law of thermodynamics: indeed, perhaps of all of physics. Part of its celebrity status stems from its reputation as the ultimate source of time-asymmetry, i.e. irreversibility, in the universe. But there are many different concepts of irreversibility, which I outline in Section 2.20. And considering these details allows me, in Section 2.21, to rebut some of the grand claims surrounding the Second Law of Thermodynamics (henceforth: TDSL). In Section 2.22, I state the Second Law in purely phenomenological thermodynamics.

In order to make progress with the project of reduction, we need to establish the crucial — and the non-essential — features of the TDSL. I highlight three features which all centre on the (different) concept(s) of irreversibility: (i) in Section 2.23.1, the importance of the environment in TD; (ii) in Section 2.23.2, quasi-static processes; and (iii) in Section 2.23.3, the distinction between the TDSL and the Minus First Law.

Then, in Section 2.24, I consider the "Demonic consequences" of the discovery of the atomic nature of matter. In particular, does the possibility of a Maxwell's demon alter the status of the TDSL? In particular, does this reduce the scope of the TDSL? Section 2.25 distinguishes four options about the scope of the TDSL.

In Section 2.26, I consider the connections between the first three of these options, 1., 2. and 3., and the connection between fluctuations and violations of the TDSL. In Section 2.27 I outline option 4: Maxwell's 'means-relative view', before suggesting an alternative: the means are determined by the lower-level theory, statistical mechanics (SM). In Section 2.28 I discuss the operation of the demon, and the SM constraints. In Section 2.29, I discuss how SM helps explain why — even if we can manipulate individual molecules — we cannot construct an engine more efficient than a Carnot engine. This explanation is known as Landauer's principle, and I explicate Wallace (2014)'s claim that this principle is 'sound' with respect to statistical mechanics but 'profound' with respect to thermodynamics; where 'sound' and 'profound' are terms of art, introduced by Earman and Norton (1998, 1999) in their criticism of Landauer's principle.

In Section 2.30, I draw together these discussions about the scope of the TDSL, and conclude that whilst one small concession must be made — the TDSL must be altered to include a 'reliability' caveat — much of the original TDSL is unaltered.

In Section 2.31, I consider what the SM realiser of the TDSL should be. Should we be seeking the Holy Grail — a non-decreasing SM function to call 'entropy'? I claim that this is neither necessary nor sufficient for a SM realiser of TDSL. The SM realiser of the TD entropy must increase in the right circumstances (namely non-quasi-static adiabatic changes) and should remain constant during quasi-static adiabatic changes. I argue that this is most naturally understood in Gibbsian SM, where Ehrenfest's principle helps us.

2.20 Concepts of Irreversibility

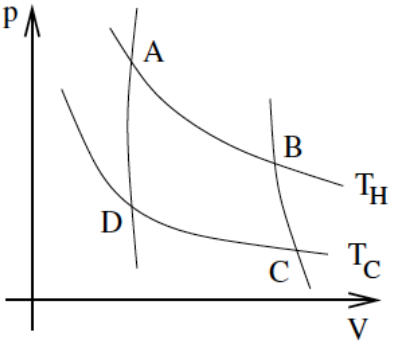

Figure 2.3: Quasi-static processes represented in the p-V plane of equilibrium states.

Before examining the TDSL, it is important to unravel the different concepts of reversibility. Uffink (2013) outlines three concepts of 'reversible' in thermal physics:

1. **Time-reversal invariance (TRI)**: there exists a map $\mathcal{T}$ — frequently assumed to be the map $t \mapsto -t$ — that maps possible histories of the system to possible histories. Call this *reversibility*$_T$.

2. **Quasi-static processes**: These ' quasi-static processes' are 'reversible', in the sense that the arrows can be drawn in either direction on the curves in Figure 2.3: corresponding to expansions and compressions. But travelling in one direction is not straightforwardly the 'time reverse' in the TRI $t \rightarrow -t$ sense: you are not performing the same interventions in a different order, but rather performing different interventions (e.g. inserting rather than removing a piston). Furthermore, quasi-static processes result from taking a limit; the limit of making very small interventions so that the system stays 'close to' equilibrium. Thus, this 'quasi-

static reversibility’ is a property of a sequence of processes, rather than of a single process (as discussed in Part 2.1.3). Call this *reversibility*$_Q$.

3. **Recoverability**: the process in question can be ‘fully undone’. The system can be returned to its initial state with no effect in the environment.[35] But the system need not retrace its steps — it can take a different path to its destination. So process P is ‘*reversible*$_R$’ i.e. recoverable, if: writing $\langle S_i, E_i\rangle \xrightarrow{P} \langle S_f, E_f\rangle$ there is a process $P*$ such that $\langle S_f, E_f\rangle \xrightarrow{P*} \langle S_i, E_i\rangle$. Planck’s terminology for such a process was ‘reversibel’ (Uffink, 2001, p. 344). [36]

2.21 The source of all asymmetry?

The celebrity status of the TDSL is illustrated by grandiose such as the following:

> The Grand Claim: the Second law is the source of all irreversible behaviour in the universe (which can be described by an increasing entropy function), and so is the naturalistic reductive base of ‘the direction of time’.

Many authors make claims that are similar to the above Grand claim. For example, (Atkins, 2007, preface) claims: “The second law is one of the all-time great laws of science, for it illuminates why anything — anything from the cooling of hot matter to the formulation of a thought — happens at all.” Such a view originates with Planck: “Every process occurring in nature proceeds in the sense which the sum of the entropies of the bodies taking part increases” (Planck 1926, p. 463). Davies (1999) considers the TDSL to be “nature’s way of driving systems towards equilibrium” (as cited in Brown and Uffink (2001)). Finally, (Hawking, 1994, p. 348) makes the following strong claim: “So the second law of thermodynamics is really a tautology. Entropy increases with time, because we define the direction of time to be that in which entropy increases”.[37]

There are two components to the Grand claim, which I will rebut in turn. (i) The TDSL is applicable to everything and it is the driving force behind all processes in nature; in this sense it is the motor of the universe. (ii) The TDSL is responsible for the arrow of time.

(i) In Chapter 4, I argue that the scope of thermodynamics is limited (because roughly, you need an equilibrium state-space Ξ and not all systems are well-represented by such

[35] Different authors vary over whether *everything* must be included in the environment, or whether some features, such as the height of a weight may be excluded. I cannot discuss this further, but see Uffink (2001) for details.

[36] Luczak (2018) adds the further condition that the process $P*$ must be one that *we* can implement. Later in this part we will see that this is part of the Maxwellian view.

[37] An admission: these quotes, with the exception of Atkins, do not establish that it is source of *all* irreversibility.

a space). Furthermore, we will see in Section 2.23.2, that there must be quasi-static processes available to complete the cycle.[38] But it does not seem, at least prima facie, that there are quasi-static processes available in Atkins' example of 'thought formation'.

The TDSL is not the reason that (most) processes happen: indeed for the type of quasi-static processes considering in the TDSL, external interventions are required to make anything happen, (as emphasised in Part 1 of this Chapter). Thus, it seems fair to say that the universe is not driven by a Carnot engine.

(ii) The second claim is that the TDSL is responsible for the arrow of time. Indeed, some go further and claim that TDSL is not only responsible for the arrow, the arrow is reducible to the asymmetry of entropy encoded in the TDSL (as seen in the above Hawking quote).

But, as we saw earlier, the type of irreversibility in the TDSL is not 'non-TRI' but rather *irrecoverability* of the initial state of both the system and the environment, especially in *non-quasi-static processes*. Thus, I agree with Uffink (2006a), that the TDSL describes the *ravages of time, rather than the arrow of time*. The loss of youth, the irrecoverability of spilt milk, the fact that Humpty Dumpty cannot be put back together again — these examples display the type of irreversibility that the TDSL describes. [39]

2.22 The Second Law Introduced

In this Section, I first state two classic formulations of the TDSL. I then discuss the idea that only a certain amount of heat can be transformed into work by explicating the Carnot cycle. As discussed in Part II of this Chapter (on the Zeroth Law), the Carnot cycle allows us to fix a temperature scale, as I explain in Section 2.22.1. But most importantly, in Section 2.22.2, we have the resources to define a new quantity: entropy.

The Kelvin Statement: "It is impossible to perform a cyclic process with no other result than that heat is absorbed from a reservoir, and work is performed" (Kelvin et al. (1882) as cited in (Uffink, 2001, p.328)).

The Clausius Statement: "It is impossible to perform a cyclic process which has no other result than that heat is absorbed from a reservoir with a low temperature and emitted into a reservoir with a higher temperature." (Clausius (1864) as cited in Uffink, 2001. p. 328).

[38] In the subsequent Section we will see that whilst the TDSL prescribes an entropy increase in processes *other* than quasi-static ones, nonetheless quasi-static processes must be *available*.

[39] Furthermore, regardless of these details, the project of 'reducing' the arrow of time to the entropic arrow faces problems, cf. Earman (1974), Price (2009).

We know the first law is: $dE = đQ + đW$. In a cyclic process, $dE = 0$. Thus, in a reversible cycle $\oint đQ = \oint đW$, which looks as if heat is being converted into work. Why doesn't this violate the Kelvin statement of the TDSL?

The crucial point is that the transfer of heat into work is not the sole effect in this reversible cycle; other things are going on. This is made clear by the Carnot cycle, which operates with two reservoirs one at T_h and the other at lower temperature T_c.

2.22.1 Carnot Cycle

There are four stages in a Carnot cycle, as shown in Figure 2.3.

1. (From A-B): the gas is isothermally expanded whilst in contact with a heat bath at T_h and heat Q_h is absorbed (i.e. the gas expands against a piston whilst in contact with a heat bath).

2. (From B-C): the thermal contact is broken and so the system is now isolated. The gas is expanded adiabatically (i.e. whilst thermally isolated).

3. (From C-D): the system is compressed isothermally at temperature T_c and heat Q_c is emitted to the heat bath.

4. Finally (From D-A): the system is isolated and compressed until it reaches its initial state A.

The net heat absorbed is the difference between the heat absorbed in the isothermal expansion, and the heat emitted in the isothermal compression: $Q_h - Q_c$. This is equal to the work done W.

The efficiency is defined by the ratio of $\eta = \frac{W}{Q_h} = \frac{Q_h - Q_c}{Q_h} = 1 - \frac{Q_c}{Q_h}$. If we could take all of the heat from the hot reservoir Q_h and turn it into work this would mean the engine had efficiency 1: but we have to give some heat, $Q_c \neq 0$, back when returning the system to its initial state.

Carnot's theorem states that the reversible cycle above is the most efficient, i.e. the best we can do, and so the ratio Q_h/Q_c is the same for all reversible engines. This is of such central importance, that sometimes presented as another version of the TDSL (Blundell and Blundell, 2009, p. 130), known as the Carnot statement:

> The Carnot Statement: No engine is more efficient than a Carnot engine: $\eta = 1 - \frac{Q_c}{Q_h}$. That is, for engine operating between two reservoirs with temperatures T_h and T_c, a reversible engine is the most efficient.

This is shown as follows. Imagine that you have two Carnot engines; one operates between two reservoirs at temperatures T_1 and T_2 (where $T_1 > T_2$) and the other between T_2 and T_3, where $T_2 > T_3$. $Q_2 = Q_1(1-\eta(T_1,T_2)$, $Q_3 = Q_2(1-\eta(T_2,T_3)$, so $Q_3 = Q_1(1-\eta(T_1,T_2))(1-\eta(T_2,T_3))$. Next, consider the two engines to be one engine, where the heat given out by the first engine, Q_2, is the heat absorbed by the second engine,Q_2. Then $Q_3 = Q_1(1-\eta(T_1,T_3)$. Then we have $1-\eta(T_1,T_3) = (1-\eta(T_1,T_2))(1-\eta(T_2,T_3))$. Because T_2 must drop out from the right hand side, $1-\eta(T_2,T_1) = \frac{f(T_2)}{f(T_1)}$. Tong (2012) says we can choose $f(T_2) = T_2$. Thus, the efficiency of a Carnot engine $\eta = 1 - \frac{T_c}{T_h}$ (cf. the two expressions for the efficiency above). We use this to define a thermodynamic temperature scale — and this (fortunately) coincides with the scale given by the ideal gas $T = pV/k_B N$.

2.22.2 Defining Entropy

The Second Law allows us to define a new state function, entropy S_{TD}. From considering the efficiency of a Carnot engine, we find $\frac{Q_h}{Q_c} = \frac{T_h}{T_c}$.

We now change notation so that Q represents the heat absorbed by the system: in the isothermal compression the heat absorbed by the system is $-Q_c$. We can also relabel as follows $Q_1 = Q_h$, $Q_2 = -Q_C$. In a Carnot cycle:

$$\Sigma_{i=1}^{2} \frac{Q_i}{T_i} = 0. \tag{2.50}$$

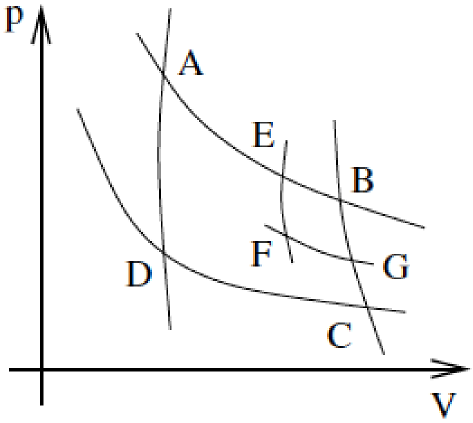

Figure 2.4: These diagram (taken from Tong (2012)) shows the original Carnot cycle, as well as another smaller Carnot cycle, EBGFE.

Now consider the reversible cycle AEFGCDA in Figure 2.4. This is the original Carnot cycle with a corner chopped out of it: the cycle EBGFE is also a Carnot cycle. We know that $Q_{AB}/T_h + Q_{CD}/T_c = 0$. Likewise in the mini Carnot cycle, $Q_{EB}/T_h + Q_{GF}/T_{FG} = 0$.

Now we can write out the heat flow in the cycle AEFGCDA:

- The heat flow in the segment FG is the reverse of GF: $Q_{FG} = -Q_{GF}$.
- Thus, the heat flow in the (non-Carnot) cycle AEFGCDA is: $Q_{AE}/T_h + Q_{FG}/T_{FG} + Q_{CD}/T_c = 0$.

By cutting more corners, i.e. by having many infinitesimal adiabats and isotherms, any reversible$_Q$ cycle in the plane can be considered. If we sum up all the contributions Q/T along the cycle, we find:

$$\oint \frac{đQ}{T} = 0. \tag{2.51}$$

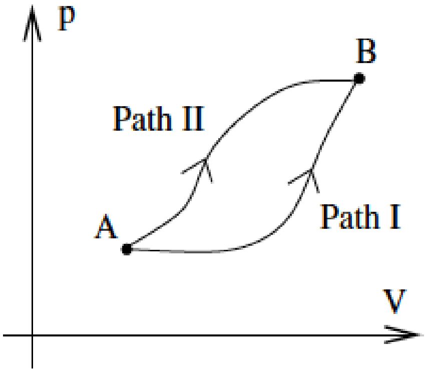

Figure 2.5: Two possible paths between two states in Ξ. Figure from Tong (2012).

Thus, if there are two (or more) reversible paths (i.e. quasi-static curves) between equilibrium state A and equilibrium state B the change in $\int_A^B \frac{đQ}{T}$ is independent of the path taken.

This (along with a reference state 0) allows us to the define a new function of state which only depends on the state variables p, V : the thermodynamic entropy S_{TD}.

$$\int_0^B \frac{đQ}{T} = S_{TD}(B) \tag{2.52}$$

Because entropy is a function of state it is path-independent: it doesn't matter how we reached state B — quasi-statically or not, or whether the system was isolated or not — either way the entropy of state B is $S(B)$.

Clausius' inequality generalises away from the reversible cycle above to any cycle:

$$\oint \frac{đQ}{T} \leqslant 0 \tag{2.53}$$

$$\oint \frac{đQ}{T} = \int_1 \frac{đQ}{T} - \int_2 \frac{đQ}{T} \leqslant 0. \tag{2.54}$$

If path 1 is an irreversible and path 2 is reversible path from state A to B, and path 1 is adiabatic (so dQ =0), then we learn that the thermodynamic entropy of an *isolated* system cannot decrease:

$$S(B) - S(A) \geqslant 0. \tag{2.55}$$

2.23 Three key features of the TDSL

In this Section I emphasis three important features of the TDSL; doing so reveals what features the SM realiser must capture, and so is important for reduction. In Section 2.23.1 I emphasis the important of the environment, in Section 2.23.2 quasi-static processes, and in Section 2.23.3 the distinction between the Minus First and the Second Law.

2.23.1 The importance of the environment

It is important to emphasis the 'sole effect' part of the Clausius statement: otherwise, fridges would be a clear counterexample to the TDSL. Fridges transport heat from a colder to hotter body — at a cost. Such transport is only prohibited as the *sole effect*. Likewise, in the previous Section, we showed that the entropy S_{TD} of the system is only non-decreasing during adiabatic (i.e. isolated) processes. Indeed, during an isothermal compression, the entropy of the system decreases. This is especially obvious when we view the Carnot cycle in the T-S plane, as shown in Figure 2.6. During the isothermal compression from C to D, the entropy of the system decreases. Of course, during an isothermal compression heat flows to the heat bath, i.e. the environment, and so during this process the *net* entropy change ΔS_{TD} is zero.

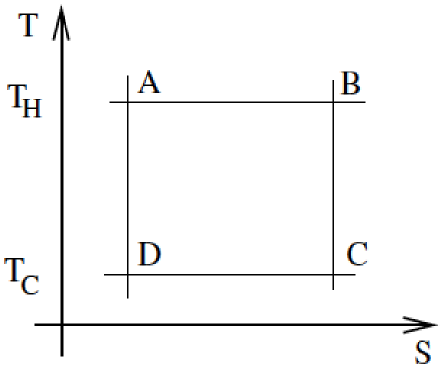

Figure 2.6: The Carnot Cycle represented in the T-S plane.

This is why considering the environment is crucial in the Second Law: we are concerned with *irrecoverability* (in Section 2.20's sense) of certain initial states during non-quasi-static processes. That is, the concept of irreversibility present in the TDSL is the impossibility of some *reversible*$_R$ processes. For some transitions, $\langle S_i, E_i \rangle \xrightarrow{P} \langle S_f, E_f \rangle$, the requisite $P*$ processes, $\langle S_f, E_f \rangle \xrightarrow{P*} \langle S_i, E_i \rangle$, to recover the initial state do not exist.

2.23.2 Quasi vs. non-quasi static processes

As discussed, only 'quasi-static processes' are represented in TD state-space. But because entropy is a state function (i.e. is path-independent), we can calculate the entropy change between two states in non-quasi processes, by considering the entropy change in a quasi-static process.

But Uffink (2006a) emphasises that such quasi-static processes must be available, and that this is not just a matter of convenience: Clausius' proof talks about *cycles* and so it requires that we can find a quasi-static process that connects the final state to the initial state in order to complete the cycle. "Indeed if such process did not exist then the entropy difference between these two states would not be defined" (Uffink, 2006a, p. 19). Whilst the existence of such a process may not be problematic in the intended applications (such as fluids), but it much less clear that this is so in e.g. living cells. "This warning that the increase of entropy is thus conditional on the existence of quasi-static transitions has been pointed out already by (Kirchhoff, 1894, p. 69)", as cited in (Uffink, 2006a, p. 19).

One might be tempted to think that whilst quasi-static processes are required during the construction of the entropy function and the above discussion of the Second law, once we have this state function we can kick away the ladder used to construct this function. As S_{TD} is a state function, we needn't have a path connecting every pair of states.[40] However, the definition of TD entropy required infinitesimal inexact differentials of heat, which is a path-dependent quantity.[41]

2.23.3 The Second Law vs. the Minus First Law

The spontaneous approach *to equilibrium* (from non-equilibrium) is distinct from the Second Law, which describes the thermodynamic entropy differences *between* equilibrium states. It is a presupposition of TD that systems *do* in fact reach a state of equilibrium. Because this requirement that systems do in fact reach equilibrium is prior to the other

[40] From discussing with Uffink: he seemed to be more concerned with the practical question... how would you know/be able to calculate the entropy difference? So need quasi-static to calculate?

[41] Whilst the existence of the integrating factor T turns the inexact differential into an exact differential (and so no longer a path-dependent quantity), we cannot 'kick away' the ladder of quasi-static processes once the entropy function has been constructed: as Uffink emphasises quasi-static processes, i.e. curves in Ξ are used to calculate the entropy S changes between different equilibrium states.

laws, Brown and Uffink (2001) call it the 'Minus First Law' (but they also suggest that it is so central that the name 'The Minus Infinite Law' would also be appropriate (Brown and Uffink, 2001, p. 529)).

The Minus First Law: An isolated system in an arbitrary initial state within a finite fixed volume will spontaneously attain a unique state of equilibrium (Brown and Uffink, 2001, p. 528).

To emphasis the contrast: the Second Law tells us that certain transitions/processes render the initial state irrecoverable, where as the Minus First Law tells us that systems spontaneously reach a state of equilibrium.

Finding the microphysical 'underpinning' for these two Laws are distinct projects (cf. Luczak (2018)). The H-theorem and coarse-graining approaches in SM are concerned with quantitatively describing the approach to equilibrium. (I will discuss my preferred account of the approach to equilibrium in Chapter 3). These foundational projects are concerned with establishing the circumstances under which a given system will approach equilibrium, rather than the quasi-static interventions on equilibrium states.

2.24 Brownian motion, the Atomic Hypothesis, and the Demonic consequences...

Imagine this caricature of history. You are a Victorian scientist, blissfully ignorant of the constitution of matter and enthralled with thermodynamics. Yet, one day, you find out the Atomic Hypothesis. Having an open minded temperament, you readily accept this hypothesis (unlike Mach). But discovering that gases are composed of molecules gives us a different description of the system according to which, violations of TDSL seem plausible. Neither CM or QM gives us reason to to think that a system cannot return to its earlier, lower entropy, state, as both are unitary.[42] And furthermore, the lower level theory, be it CM or QM, is more fundamental — thus casting doubt on the TDSL. Should we still believe the TDSL to be true? (Of course, history was quite the reverse of my caricature: the epistemic security of TD led many to doubt the atomic hypothesis, rather than vice versa). Because there is nothing in CM/QM that suggests that (isolated) systems cannot return to earlier lower entropy states, the TDSL is not true on the basis of molecular dynamics alone — it is not solely a consequence of CM/QM.

But this is no surprise. Recall that there are three different levels of description under consideration: (I) TD, (II) SM and (III) CM/QM. Thus, it is unsurprising that the TDSL is not true solely on the basis of QM/CM. It has been appreciated since

[42] Modulo worries about the measurement problem of course.

Boltzmann that additional assumptions, e.g. a Past Hypothesis, are required. But these are the ingredients of statistical mechanics. Thus, these additional assumptions are only indirectly connected to thermodynamics, since SM is the stepping stone which aims to build the laws of thermodynamics from mechanics.

In this Section I discuss the implications of the atomic nature of matter for the TDSL: is the TDSL still true in light of the these lower-level theories? Brownian motion played an important role in the discovery of the atomic nature of matter — but it also threw a shadow of doubt over the TDSL: as I discuss in Section 2.24.1. The possibilities of violations of the TDSL are made especially vivid by two demons, which I describe in Section 2.24.2 and 2.24.3. Describing the doubt cast over the TDSL by these demons prepares the ground for Section 2.25, where I consider whether the scope of the TDSL needs to be limited.

2.24.1 Brownian motion

Originally discovered by Robert Brown in 1827, the phenomenon of Brownian motion can be easily viewed through a microscope. Pollen grains suspended in a solution can be seen to be jiggling around, undertaking what is often referred to as a random walk. The suggested explanations were diverse: Brown initially thought that the phenomenon was intimately connected to life, and W. Stanley Jevons thought it was connected to osmosis. But later in the 19th century the phenomenon was connected to kinetic theory: for example, Cantoni in 1868 claimed that Brownian motion is a "beautiful and direct experimental demonstration of the fundamental principles of the mechanical theory of heat" (Cercignani, 1998, p. 215).

But contra Cantoni, many physicists, such as the French Physicist Leon Gouy, claimed that Brownian motion would violate the TDSL. "This led to an important comment by Poincare in an address at the congress of St Louis (1904): 'if this be so, to see the world return backward, we no longer have need of the infinitely subtle eye of Maxwell's demon; our microscope suffices us' "(Cercignani, 1998, p. 215).

Einstein's 1905 paper was central in cementing the idea that Brownian motion is problematic for the TDSL (Cercignani, 1998, p. 216). Popper is particularly strident in this respect: "the entropy law, in Planck's formulation, is simply falsified by the Brownian movement, as interpreted by Einstein" (Popper, 1957, p. 152).[43]

Why do these authors think that Brownian motion is a violation of the TDSL? The Brownian particle (i.e. the pollen grain) is much larger than the molecules it is sur-

[43] The Planckian formulation of the TDSL that Popper quotes is: " There does not exist a perpetual motion machine, of the second order, that is to say, a physical system, immersed in a heat bath, which by cooling down (or, which is the same, by emitting heat to the surrounding heat bath), can move a heavy body against a force, thus increasing its potential energy; or in terrestrial terms, a machine which by cooling down, can lift a weight" (Popper, 1957, p.151).

rounded by. The random collisions between these molecules and the Brownian particle do not cancel out. Instead, the Brownian particle travels $\sim \sqrt{N}$ steps for every N collisions. The random motion of the water molecules — construed as heat — can be converted in to work — construed as the macroscopic motion of the Brownian particle. There has been no other effect: and so this looks like a prima facie violation of the Kelvin statement. But whilst heat has been turned into work with no other effect, it is unclear whether this is a cyclic process in the manner than Kelvin or Clausius requires.

Another, related, way of seeing how the microscopic nature of thermal systems causes problems: there are random thermal fluctuations and these can allow violations of the TDSL — albeit small ones. For example, there could be a fluctuation such that a small amount of energy was transferred from a colder to a hotter system with no other effect: contra the Clausius statement of the second law. But once again it is unclear that this qualifies as a cyclic processes.

Next I consider two demons, whose ability to violate the TDSL is not in doubt.

2.24.2 Loschmidt's demon

Loschmidt's demon will be discussed in detail in Chapter 3, when considering the reversibility objection. This objection was originally raised against Boltzmann's H-theorem, but it also applies to other frameworks in non-equilibrium SM for describing the approach to equilibrium such as the coarse-graining framework that I will discuss in Chapter 3.

The heart of the objection is that the lower-level theory, be it QM or CM, allows a system to retrace its steps if a suitable intervention — such as reversing the momenta of all the gas molecules — is allowed. Indeed, in the Spin Echo experiment, a spin system can be forced to retrace its steps if the B-field is changed in a particular way (cf. Chapter 3 and Hahn (1950)). More generally, the possibility countenanced by CM for the system to retrace its steps was seen as problematic for the H-theorem (and similar projects) as it suggests that 'anti-equilibration' processes are possible — contra the Boltzmannian equation and other equations in non-equilibrium SM. Justifying why these 'anti-equilibrium' processes are not something we experience, or are at least the exception not the rule, is a core philosophical project in non-equilibrium SM.

But to some extent, Boltzmann's original reply to Loschmidt, hits the nail on the head. Boltzmann is said to have retorted to Loschmidt 'go ahead then, reverse the momenta'! His challenge of course rests on the fact that it is hard to see how we could implement such an intervention, given our clumsy grip on the world. A 'Loschmidt demon' on the other hand would — ex hypothesis — be able to implement such an intervention.

In Chapter 3, I will argue that Loschmidt's objection reveals that the time-asymmetry in SM is a feature of our perspective. If we could readily reverse the momenta we could reverse the trajectories and see 'anti-equilibration'. Had we humans been more nimble, Loschmidt could have reversed the momenta in response to Boltzmann.

Loschmidt's demon is a problem for the construction of SM from CM/QM. But we can also see how it could cause problems for the TDSL: a Loschmidt demon can 'reverse the momenta' of all the elements of the system — thus forcing the system to retrace its steps (to its earlier low entropy state).

An example of a Loschmidt demon: do an isothermal expansion (thus, transforming heat to work) then press the 'restore button'. This 'restore button' acts like a Loschmidt demon and reverses the momenta in such a way as to return the system to its initial state. But before this Loschmidt demon can claim they have violated the TDSL, they must consider their actions. We must account for how the restore button works because the TDSL includes the environment: does *its* entropy increase? (As I emphasised when considering the essential features of the TDSL, thermodynamic entropy decreasing is only interesting from a foundational viewpoint if the system is isolated).

2.24.3 Maxwell's demon

Unlike Loschmidt's demon, Maxwell's demon has the TDSL squarely in its line of fire. The demon is a nimble-fingered creature, which operates a trapdoor in a partition between two gases (which are initially in the same macrostate) — and only lets the fast molecules through from the right side to the left. Likewise, the demon only lets slow (i.e. below average velocity) molecules from the left to the right. This leads to the left hand side becoming hotter than the right hand side. Thus, the demon has transferred heat from a colder to a hotter body with no other effect — a violation of the Clausius statement of the TDSL.

Whilst Maxwell's original demon involved an intelligent creature (the '*intelligent demon*'), there are now a range of Maxwellian demons. Following Szilard (1929) it has become popular to discuss a one-molecule gas. Initially the one-molecule gas can explore the entire volume V of a container. A partition is then inserted at the centre, so the gas now occupies $\frac{V}{2}$ of the container. A measurement is made by the demon to see which side of the partition the gas is on. A piston is then inserted through the sub volume not containing the particle, and the gas is then isothermally expanded back to its initial volume, by the particle pushing the piston. There has been a cyclic process whose sole result is to have turned heat (from the heat bath) into work (by pushing the piston), as depicted in Figure 2.7. Call this the *miniature demon* (not because the demon itself must be small, but because it operates on small systems).

To connect this to Section 2.20, recall that P is '*reversible$_R$*' i.e. recoverable if given a process $\langle S_i, E_i \rangle \xrightarrow{P} \langle S_f, E_f \rangle$ there is a process $P*$ $\langle S_f, E_f \rangle \xrightarrow{P*} \langle S_i, E_i \rangle$; whilst the TDSL says there exists some processes I which are not *reversible$_R$*. These processes I take

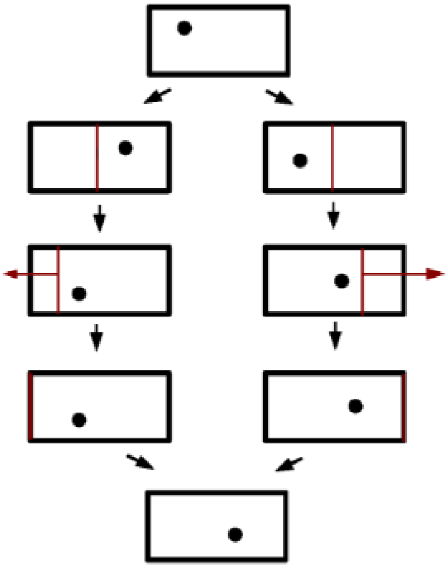

Figure 2.7: The operation of the miniature demon: the demon inserts a piston depending on which side of the partition that gas was trapped in. The expansion at the end is done whilst the gas is in contact with a heat bath, so that there is a cycle whose sole effect is the conversion of heat to work.

the $\langle S_i, E_i \rangle \xrightarrow{I} \langle S_f, E_f \rangle$, where the thermodynamic entropy S_{TD} associated to $\langle S_f, E_f \rangle$ is greater than the thermodynamic entropy associated to $\langle S_i, E_i \rangle$. Because processes such as I are irrecoverable there is no process $I*$ which takes the system and environment back to its initial state. But Maxwell's demon suggests that there is a process $I*$ that takes the above system back to its initial state: the demon can implement this process.

Crucially, however, we have to consider whether the state of the demon has an entropy change associated to it. Because it is the entire system (i.e. system, plus demonic operations) that, according to the TDSL, must be (thermodynamic) entropy non-decreasing. That is: $I*$ must not only take $S_f \rightarrow S_i$, it must also take $E_f \rightarrow E_i$, where the demon is included in the environment.

Thus, the key caveat is: to establish that these Maxwellian, or Loschmidtian Demons violate TDSL, we must consider the state of the environment — i.e. how the demons operate. Yet we can see that there a certain unity to these challenges to the TDSL: they all stem from the atomic nature of matter. Whether Brownian motion counts as an example of a violation of the TDSL is debatable. But all agree that if it is, it is only a *microscopic* violation: no one hopes to harness these 'violations' to build an engine more efficient than a Carnot engine.

Yet a Maxwellian demon can harness these microscopic fluctuations to create bona fide violations of the TDSL. Indeed, Norton (2011) argues that the demon operates by scaling up such microscopic fluctuations into full blown macroscopic violations. This raises the question: what implications do these demons have for the truth of the TDSL?

2.25 Four different views about the scope of the TDSL

What should we make of Brownian motion and these demons? Do they show the TDSL to be false? Yet the TDSL does seem to capture something true about the world: engines more efficient than a Carnot engine are hardly a dime a dozen. Rather than out-and-out falsity, these considerations have been suggested to alter the scope of the TDSL — so that it is no longer strictly true.

The idea that the TDSL is not a strict law was suggested by Maxwell: "Hence the SL of TD is continually being violated, and that to a considerable extent, in any sufficiently small group of molecules belonging to a real body. As the number of molecules in the group is increased, the deviations from the mean of the whole becomes smaller and less frequent; and when the number is increased till the group includes a sensible portion of the body, the probability of a measureable variation from the mean occurring in a finite number of years becomes so small that it may be regarded as practically an impossibility" Maxwell (1891) as quoted by Cercignani (1998).

Thus Myrvold (2011) distinguishes four different possible views of the scope of the TDSL[44] :

1. *(The Strict view)*: there are never violations.
2. *(The Probabilistic view)*: it is very unlikely that there will be large violations. (Fluctuations motivate this view).
3. *(The Statistical view)*: the TDSL only applies to large numbers of degrees of freedom (here after DOF). (Brownian motion motivates this view — as do Boltzmannian considerations, as we will see below).
4. *(The Maxwellian view)*: the TDSL only applies to large numbers of DOF and is contingent on our current, but perhaps temporary, technological inability to manipulate individual molecules. (Maxwell's demon motivates this view).

One crucial difference between 1.-3. and 4. is that, according to 4., the TDSL *could* be flat-out false — if we found a Maxwell demon: which in the future, we may.

Considering the scope of the TDSL is part and parcel of the reductive project — rather than prior to it, contra Myrvold.[45] The demons and the difficulties for the TDSL arise from the depths of the lower-level theories about the molecular nature of matter. As

[44] Myrvold (2011) raises the Maxwellian view as deserving consideration, but does not explicitly endorse it.

[45] Myrvold says we should think about which version of the TDSL is correct because for the project of reduction it matters *which* version of the TDSL we are trying to recover.

such, questions about alterations to the TDSL cannot be disentangled from questions about reduction. This is unsurprising given Chapter 1's discussion of how reduction might limit the domain of applicability, i.e. scope, of the higher-level theory, T_t.

In considering the scope of TD we will use resources from outside TD. Considering fluctuations — and how the probabilistic and statistical views of the TDSL coincide — involves studying the different ensembles of SM. Thus, some of the discussion will stray from the original TD consideration of Carnot engines to the SM considerations of ensembles. Indeed, sometimes it is the image of the TDSL in SM, rather than the TDSL itself, which some authors have in mind when discussing the scope of the 'TDSL'.

To adjudicate between options 1.-4., I will proceed in two stages. First, in Section 2.26, I discuss the way in which options 1.-3. come together: in the thermodynamic limit they coincide. This is because fluctuations vanish in this limit. Considering the connection between fluctuations and violations of the TDSL requires us to discuss Maxwell's demon, who exploits these fluctuations to create violations.

Thus, the second stage of this project is to consider Maxwell's demon and what implications it has for the status of the TDSL. In Section 2.27 I resist (parts of) the Maxwellian means-relative view of the TDSL; namely that there exists no physical principle which can rule out such a demon. This launches an extended discussion, in Section 2.28, of Maxwell's demon. In Section 2.29, I argue that SM provides such a physical principle: Landauer's principle.

Finally, in Section 2.30, I bring these two stages together to conclude that alteration of the TDSL in light of the demons and difficulties is only slight: a caveat about the reliability of Carnot engines must be included. Thus, I will endorse a version option 2. — but suggest that this is very close to option 1.

2.26 Connections between options 1.-3.

In this Section, I discuss the connections between the size of the system and the fluctuations that motivate option *(2. The probabilistic view)*. But why think that fluctuations lead to violations of the TDSL? I discuss one bad reason for downgrading the scope of the TDSL to options 2. or 3, which involves mistaking the TDSL for its image in the Boltzmannian framework. I will then discuss how fluctuations lead to violations with the help of a demon.

From our modern perspective, the distinction between *(2. The probabilistic view)* and *(3. The statistical view)* is often blurred. But as Myrvold emphasises, for Maxwell, 'statistical' did not mean 'stochastic' and so 3. is strictly weaker than 2. Of course, the two are connected: as the size of the system increases, then the probability of fluctuations decreases.

This can be seen as follows. The thermodynamic limit is, roughly, the limit where the number of particles in the system tends to infinity: $N \to \infty$. In the thermodynamic limit, the microcanonical and canonical ensembles coincide. This is because the probabilistic fluctuations scale with N. This can be seen as follows. The variance in energy,

$$\Delta E^2 = \langle (E - \langle E \rangle)^2 \rangle = \langle E^2 \rangle - \langle E \rangle^2 \tag{2.56}$$

can be written in terms of the partition function: $\Delta E^2 = \frac{\partial^2}{\partial \beta^2} log Z = -\frac{\partial \langle E \rangle}{\partial \beta}$. There is another way to express these fluctuations. In the canonical ensemble, the definition of heat capacity is: $C_V = \frac{\partial \langle E \rangle}{\partial T}\Big|_V$. Since, $\beta = 1/k_B T$, the spread of energies can be written:

$$\Delta E^2 = k_B T^2 C_V \tag{2.57}$$

ΔE describes how the energy of the system fluctuates in a probabilistic manner, while C_V describes the ability of the system to absorb energy. Equation 2.57 is an example of the fluctuation-dissipation theorem. It tells us that a system with a high heat capacity (i.e. a system that can absorb a lot of energy without changing its temperature very much) will have more fluctuations than a system with a low heat capacity.

The above equations can also be used to demonstrate how the fluctuations scale with the size of the system. Typically $E \sim N$ and $C_V \sim N$, and this means that:

$$\frac{\Delta E}{E} \sim \frac{1}{\sqrt{N}}. \tag{2.58}$$

Thus, in the thermodynamic limit, the energy probability distribution becomes more and more peaked around the mean value $\langle E \rangle$ so that energy can be treated as essentially fixed, as is the case in the microcanonical ensemble. The probability of fluctuations decreases as the size of the system increases — and this is why from our modern perspective options 2. and 3. are often run together. And in the limit of an infinite number of components, there are no fluctuations. Thus, in the thermodynamic limit, (1), (2) and (3) coincide.

Yet this raises the question: why do fluctuations lead to violations of the TDSL? Is a fluctuation *on its own* sufficient to violate the TDSL — so that it is downgraded to being a either a probabilistic or statistical truth? There are two arguments that it is downgraded: one bad, one good.

The bad argument is as follows. Thus far I have discussed this issue in a Gibbsian framework. But whilst the Boltzmannian framework is not the focus here, there is one issue which is so central that it is worth discussing. The idea that the TDSL is a 'statistical truth' has its roots in the Boltzmannian framework. According to the Boltzmannian framework, the system can fluctuate away from the equilibrium macrostate, and in doing so the Boltzmann entropy, S_B, decreases. However, the larger the system the more infrequent these fluctuations away from equilibrium will be. Thus, 'the Boltzmann entropy is non-decreasing' is only a statistical truth.

But notice: the Boltzmann entropy, S_B, is a SM concept, and so unless the S_{TD} is *identified* with S_B, it is not immediate what the consequences are for the scope of the TDSL. Instead, a putative image of the *TDSL at the SM level* is only a statistical truth. I will set this aside because: (a) I will later discuss why a decreasing SM entropy need not always have implications for TDSL, when I consider the realiser of the TDSL in SM in Section 2.31; and (b) as stated earlier, I am going to focus on the Gibbsian, rather than Boltzmannian framework.

The good argument is as follows. Fluctuations are central to considering violations of the TDSL because, as Norton (2011) says, a demon uses fluctuations in their schemes to create violations of the TDSL. The fluctuations can happen in a random direction. In the case of the one-molecule gas, the volume is likely to spontaneously decrease to the right or to the left hand side. As these changes are fluctuations, they are unpredictable. But the miniature demon finds out the system's state, and acts according to create a cycle that violates the TDSL.

But there is an 'ignorant. demon: call this is the *lucky demon* (if indeed he is worthy of the name 'demon'). The direction of the fluctuations cannot be predicted, and so the demon cannot predict which side to insert the piston in order to then be able to isothermally expand the gas back to its original state (thus turning heat entirely into work). This lucky demon is ignorant, unlike the miniature demon: it doesn't find out which side of the partition the gas is on. Instead this lucky demon just guesses — and gets it right! Thus, even if this is a one-time occurrence, we have still got one cycle more efficient than a Carnot cycle. Of course, this lucky demon is hardly *reliable* and so cannot be used to consistently violate the TDSL.

Yet, even this unreliable demon would not be able to operate on an infinite system — as in such a system, there are no fluctuations. Many of the macroscopic systems that TD considers are 'approximately infinite'. To get a sense how unlikely fluctuations are for systems of the order 10^{23}, Tong (2012, p. 8) points out that the timescales over which we would expect to find the system with an energy other than the equilibrium value, are of the order of exponentials of exponentials. Such numbers are so large that it makes little difference which system of units we use.[46] Hence the lucky demon does not seem to pose a severe threat to the scope of the TDSL.

But the intelligent demon and the miniature demon, on the other hand, use their knowledge of the system's state in order to reliably and consistently violate the TDSL.

[46] Tong (2012) says that number of microstates for a two-state system of 10^{23} components as $2^{10^{23}}$ which is a number so large, that imagining it as a distance — it would not matter whether you measure it in microns or lightyears.

The importance of the fluctuations — and so the assessment of options 2. and 3. for the scope of the TDSL — depends on whether a demon can harness these fluctuations to create violations. It is to these fully fledged demons to which I now turn.

2.27 The Demons and the Maxwellian means-relative view

Loschmidt's demon reveals that the time-asymmetry in SM is a feature of our perspective. If we could readily reverse the momenta we could reverse the trajectories and see 'anti-equilibration'. Is Maxwell's demon like Loschmidt's demon? Maroney (2009) suggests that Maxwell thought so, he says: "The operation of Maxwell's demon is simply a matter of scale and the statistical nature of the second law is not probabilistic, but due to our inability to discriminate the exact state of a large number of particles (similar to our inability to exploit Loschmidt's reversibility objection)" (Maroney, 2009, §1.1).

If we were able to manipulate individual molecules, would we expect to see violations of the Second Law? Maxwell thought so: "For Maxwell, no matter of physical principle precludes the operation of a Maxwell demon; it is only our current, but perhaps temporary inability to manipulate molecules individually that prevents us from doing what the demon would be able to do" (Myrvold, 2011, p. 2).

Because we can only manipulate 'large aggregates' rather than individual molecules, we cannot perform demonic manipulations. Thus, according to the Maxwellian view, the validity of the TDSL is limited to these 'large aggregates', i.e. macroscopic realm. As such, Maxwell held that the TDSL was only applicable to systems with large no. of DOF. Myrvold (2011): "if there were an agent capable of manipulating individual molecules, then according to Maxwell, the distinction between heat and work would break down". According to this Maxwellian view, we should never have expected the TDSL to apply to a one-molecule gas: it is outside the scope of the TDSL.

To sum up: Myrvold ascribes a 'means-relative' view of thermodynamics to Maxwell. If our means were to change, then we might be able to engineer violations of the SL, thus decreasing its scope even further — in order to exclude these situations. (To continue the analogy with Loschmidt's demon: later physics revealed that we can sometimes reverse the momenta — in the spin echo experiment). If the TDSL just stems from *our* inability to manipulate microscopic systems, then the irrecoverability of certain states is just down to *us*. A process $I*$ is available after all, but not to us — only to a Maxwell demon. Myrvold (2011) suggests that if the TDSL stems from our inability, then perhaps anthropomorphising the demon might not be a mistake, contra Earman and Norton (1998, 1999, p. 4).

2.27.1 The difference between Maxwell's demon and Loschmidt's demon: resisting the Maxwellian view

I motivated the Maxwellian view by claiming that Maxwell's demon had the same status as Loschmidt's demon. But I think these demons differ in an important way: and this undermines the Maxwellian view. (But much of the spirit of the Maxwellian view — minus the anthropocentrism — is found in the control theory view of thermodynamics as explicated in Part 1 of this Chapter).

Maxwell claimed there was no 'physical principle' that there could be no Maxwell demon — and this is much like Loschmidt's demon. Indeed, there is nothing in physics that rules out a Loschmidt demon (indeed, the spin echo experiment is an example). This is because there are no physical theories to consider between the level of CM and the level of SM, which could supply such a 'physical principle'.[47] But unlike Loschmidt's demon, there is a physical principle why we cannot have a Maxwell demon. Or weaker: there's a physical theory at an intermediate level of description we can consider to see whether it is possible to have an Maxwellian demon: statistical mechanics.

The Maxwellian view claims that whether certain processes (i.e. $I*$) are possible, depends on the means available. Of course, the 'means' are constrained by our epistemic and cognitive limitations — whether we are too clumsy to manipulate individual molecules. (An example of an epistemic limitation: if the two gases on either side of partition are different, but if you don't know this then you can't exploit this fact to do work. Indeed, this phenomena lies behind the entropy of mixing, cf. Gibbs (1878).[48] But ultimately, our means are also constrained *physically*: i.e. by the lower-level theory. According to the CM/QM level, there seems no reason to suggest that a Maxwell demon is impossible. But our means might also be constrained by the intermediate theory: statistical mechanics. The question is whether this theory can provide a 'physical principle' to rule out a Maxwellian demon — and thus suggest that technological advances will never allow us to violate the TDSL.

Maxwell conceived of the demon in 1867 - over 150 years later, it is common practice to think of manipulating single molecules (cf. single shot quantum thermodynamics, as discussed inter alia Horodecki and Oppenheim (2013a), Del Rio et al. (2011), Brandao et al. (2013)). But, in the next Section, by considering SM we will see that having the ability to manipulate individual molecules will not mean that we can violate TDSL —

[47] As I will discuss in Chapter 3, the 'non-SM' behaviour of the Loschmidt demon is due to special initial conditions, but there is no principled way to rule out such initial conditions. According to lower-level theories of QM/CM, these initial conditions look like all other initial conditions and are not special. Yet SM itself doesn't have the resources to explain why these initial conditions are problematic, since the displayed behaviour is outside the jurisdiction of SM.

[48] Whilst the Gibbs paradox is another, distinct motivation for the Maxwellian view, I cannot discuss it further here.

and so we can't construct an engine more efficient than a Carnot engine. (Of course, as we will see in Section 2.29.3 this involves some weak assumptions within SM). Indeed this conclusion should be unsurprising: if it were just our clumsy fingers that were getting in the way of an engine more efficient than a Carnot engine, it would be hive of research.

2.28 Statistical Mechanics and Maxwell's demon

In this Section I first consider what it is means to talk of physical possibility — and why thermodynamics does not have the resources to remove the demon. In Section 2.28.2, I then consider what is at stake with the demon, and what it is that needs to be explained. In Section 2.28.3 I discuss the key component required for an Maxwell demon to operate: controlled operations (also known as feedback). Performing certain interventions on the system that depend on the system's state, generically allow violations of the TDSL. But I discuss how we can analyse feedback processes as physical — within *statistical mechanics*.

2.28.1 Thermodynamics and possibility

Is a Maxwell demon physically possible? Whether something is possible depends on the constraints on 'possibility'. No doubt there are metaphysically possible worlds where Maxwellian demons dwell. But are there nomologically possible worlds containing such demons? It depends on which theory's laws we consider: whether we consider SM or TD — or another theory.

According to TD, for example, a Maxwell demon is not possible. According to TD, if the demon is bounded by the TDSL, then there must be a compensating thermodynamic entropy increase somewhere. But why think that a Maxwellian demon obeys the laws of TD? Plenty of much less exotic systems do not obey TD, as I will discuss in Chapter 4. Given my complaints about the narrow scope of TD, the assumption that a Demon/any putative demon should obey the SL is unwarranted. At the very least an argument needs to be given that the TDSL applies to everything, in order for this strategy for considering whether a Maxwell demon is possible to succeed.

A better strategy for stopping the demon is to point out the importance of the environment in thermodynamics. The entropy of a system decreases in an isothermal compression, but this is not a violation of the TDSL as heat flows to the environment, whose entropy thus increases. Is the demon only an *apparent* violation, because we have failed to take into account the state of the environment?

But the demon is disanalogous to the isothermal compression example, since —prima facie— the demon does not need to be in thermal contact with the system to operate.

Just assuming that TD is applicable to the demon and that there will be the appropriate heat flow so that that TDSL is not violated seems like an unsatisfactory tactic without some explanation of how the demon works. Clearly, we need to consider the demons operation in more detail (as will be done in Section 2.28.3). But first I consider: what is the explanandum about Maxwell's demon at stake?

2.28.2 The explanandum

What is the aim of considering Maxwell's demon? In what follows I do not 'aim to save the TDSL'. Such an aim is the motivation in much of the literature — and a similar dialectical move is made in the controversial realm of black holes: the 'Generalised second law' is sometimes claimed to be needed in order to 'save the ordinary SL' (cf. Bekenstein (1973)). In Section 2.21, I deflated some the grand claims surrounding the TDSL. If the TDSL does not have the grand status that Planck and others attributed to it, why worry if it is violated? After all, it would be much more interesting if the TDSL *were* violated and we could solve the world's energy crisis.

Yet the TDSL seems to capture something true about the world. Thus, the aim of my discussion of here is explain how the TDSL can capture something true, in the face of the demons. Of course, if we cannot find such an explanation, then it may be that the Maxwellian view is right: the TDSL is a contingency based on our current lack of technological ingenuity. We have already seen that TD itself does not provide much satisfying insight into the possibility of a Maxwell demon. According to QM or CM, the existence of such a demon looks plausible. But in order to go beyond mere plausibility and to consider whether the demon is possible according to the lower-level theory, we need to translate TD into CM/QM — which practically speaking means looking at SM. Thus, my aim is to consider whether a Maxwell demon is possible according to SM.

2.28.3 Details of the demon: controlled operations

I think there is a feature common to all Demons: how the demon acts, i.e. which interventions it performs on the system, depends on the state of the system. For Maxwell's 'intelligent' demon, whether a molecule was let through the trapdoor depends on the state of the molecule, i.e. how fast it is moving. For the 'miniature demon' (that operates on a one-molecule gas turning heat-solely-to-work), which side of the partition the piston was inserted, depends on the state of the gas: was it trapped on the lefthand or righthand side?

Sometimes this 'how the demon acts depends on the state of the system' is called a controlled operation, since the state of the system 'controls' which processes happens. As a general definition, a controlled operation is as follows: if the control system is detected to be in state X, then perform process p_X on the target system Z. If the control state is detected to be Y, then perform process p_Y on Z. Another word for this type of process is 'a feedback' process.[49]

Knowing the state of the system is crucial here: controlled operations massively increase the states the target system Z can be transitioned into. Norton (2005, 2011) gives lots of examples where, once the state is known, the possibilities of manipulations seem unlimited. For example, if we know which side of the box the one-molecule gas is on, we can frictionlessly swivel the box so that a box with a gas on the RHS becomes a box with the gas on the LHS (or vice versa). Thus I claim: feedback or control operations are essential to a Maxwell demon that assuredly violated the TDSL.

Now, we need to consider the situation from the perspective of SM. The idea that feedback processes are key can be explicated formally, following Wallace.

Take a system, such as a gas in box in thermal equilibrium. This system can be given a SM description: $\rho_{cg}(V_1)$, and a SM entropy, S_G. Different interventions can then be performed on the system. If these interventions are adiabatic or isothermal changes to external parameters, then it can be shown that the system can be transitions into any state $\rho_{cg}(V_2)$, provided that $\Delta S_G \geqslant 0$.[50] That is, these isothermal /adiabatic interventions can alter the system's state, but there is a constraint: S_G cannot decrease (cf. (Wallace, 2014, p. 704)).

But if the toolbox of interventions is expanded to include feedback processes, then the system can be transitioned into any state — in particular, it can be transitioned into a state with lower S_G. Hence, Wallace's control theory with feedback shows in a very general way that 'feedback' massively increases our control of the system.

If the 'means' are the interventions that we have access to, then there is a clear sense in which the means relative approach ties in here. Here there is something that the 'means-relative' view of thermodynamics gets right. The means-relative view is right in that we can clearly show that expanding the range of possible interventions on the systems — the 'means' — to include feedback (controlled operations), this drastically expands the possible transitions the system can be forced to undergo. In particular: it seems that the TDSL can be assuredly violated.

But I believe that the means-relative view is wrong, since there are 'in principle' constraints on the available means — contra Maxwell. Maxwell leaves the possibility of a Maxwell demon open to future technology. But there are some constraints placed by the lower-level theory, which I now consider.

The first, and key, assumption is that the system (which could be a demon) imple-

[49] In Wallace's terminology, 'control process' refers to any intervention on the system, and 'feedback process' refers to a control operation in the sense above.

[50] How S_G relates to S_{TD} will be considered later in Section 2.31. Note that Wallace's demonstration that isothermal and adiabatic interventions cannot transition the system to a state with a lower S_G, requires 'parameter stability': roughly, a condition that systems reach equilibrium.

menting the feedback process, or controlled option, is itself "physically analysable" (Wallace, 2014, p. 719). There can be no 'deus ex machine' in Feyerabend (1966)'s words — the controlling system must be able to be treated as a physical system. We have the controlled system, the controlling mechanism, and the usual array of heat baths, pistons and partitions. If the controlling mechanism is also a physical system, then we can draw a line around it and the controlled system and treat them together, as one large 'thermal object' — which can be intervened on – but without feedback/controlled operations. This 'automation constraint' means that the demon cannot be a god outside the system, whose intelligence is essential to its tricks — instead, the demon is treated like a computer. The controlling system performs different operations depending on the state of the system, but the mechanism to decide which processes is implement must be 'includable' in the whole system.[51]

Now, for the second assumption — and this is where is SM provides an insight. Having included the feedback mechanism into the system, we now have one large system, which does not have a controlling system, i.e. *without* feedback processes. This system is a no-feedback system, the only interventions are the familiar adiabatic and isothermal interventions. And so we know for this total system, S_G cannot decrease. Thus, if the controlled sub-system's SM entropy S_G decreases, there must be a compensating increase in the controlling sub-system. Here the part of SM that we are invoking is the preservation of phase space volume: the underlying microdynamics of CM/QM are unitary, and so the phase space volume associated to ρ is constant.

But the question now is: how does such a compensating increase come about? Why should we think that S_G of the controlling mechanism should increase? And here is where Landauer's principle comes in.

2.29 Landauer's principle

Roughly speaking, Landauer's principle (LP) states that there is an entropy increase of $S_G = k_B ln 2$ associated to resetting one bit of data. Why are we considering resetting? As above, the demon implements a 'controlled operation' and consequently, as I explain in Section 2.29.1, this requires the demon to have a 'memory'. In Section 2.29.2 , I motivate why the demon needs to *reset* its memory. In subsection 2.29.3 I sketch the quantitive content of Landauer's principle.

[51] In the work of Ladyman et al. (2008, 2007), this idea is captured by an 'L-machine', but I cannot discuss this further here.

2.29.1 Why the demon needs a memory

The demon is going to perform a *controlled operation* — the process implemented on the target system will depend on the the state of the control system. So the state of the system has to be measured and stored.

In abstract terms, a controlled operation is a logical transformation that maps an input state of at least 2 bits to an output state of at least two bits in such a way that how one of the bits, the target bit, is transformed depends on the value of the other bit, the control bit. The most commonly discussed example of such an operation is CNOT which is defined by the following table, where bit_1 is the target bit and bit_2 is the control bit.

$Input_1$	$Input_2$	$Output_1$	$Output_2$
0	0	0	0
1	0	1	0
0	1	1	1
1	1	0	1

But a crucial point is that the control bit must be represented by a distinct degree of freedom from the target bit. For physical degrees of freedom to represent different bits they must be independent of each other. This is because, for a physical system to represent a set of n bits it must have sufficient, i.e. 2^n, different configurations of its degrees of freedom. For example, if the only two alternative states for a one molecule gas are 'B_L': the molecule trapped on the LHS, or 'B_R': the molecule is trapped on RHS, then it does not make sense to say this represents 001010, since this string represents 6 bits of information for which the system would need at least 2^6 different configurations. If this were not so the different bits could not vary independently of each other.

The demon will implement different processes/interventions depending on the system's state. The system (i.e. the one-molecule gas) cannot be the target system *and* the control system: Ladyman et al. (2007) are explicit that "the same bit cannot be both the control and the target of a controlled operation" (Ladyman et al., 2007, p. 23).[52]

There is another reason why different bits cannot be represented by the same physical state at different times, namely that to allow them to do so completely trivialises the physical implementation of logical transformations. For example, consider the logical transformation, COPY. If relabelling is allowed, then no physical change in a system is

[52] This point is considered further in Ladyman and Robertson (2013). The key clarification about controlled processes, in that paper, serves to show why Norton (2011)'s objections to Ladyman et al. (2007) fail. The heart of Norton's objections lies in the fact that he thinks that a controlled operation can be implemented by a physical system that has only one degree of freedom: which must therefore be used to represent both the target bit and the control bit. But, as above, Ladyman et al. (2007) are explicit that "the same bit cannot be both the control and the target of a controlled operation".

required for it to implement COPY: because we can simply stipulate that whatever state the physical degree of freedom happens to be in now represents the state being copied.

For example, the four different rows of the CNOT table could be physically represented by a pin on four different places on a chess board (here we do not have independent degrees of freedom). Since there are two different positions corresponding to the single value of the control bit, which physical process acts on the system depends on the value of both bits, not just the control bit. Of course, the physical implementation of logical processes is a vast and fascinating topic, but I cannot delve any further into it here.

The key point is: the demon needs to have a memory, which stores the state of the target system, and controls which process is implemented on that system.

2.29.2 Why the demon needs to reset its memory

Often it is claimed that the demon needs to reset its memory so that the process is truly *cyclic*. But why should we care that the process is cyclic? Agreed, the original formulations of the TDSL talk of cycles, but if we can make an engine more efficient than Carnot - and solve the energy crisis - why worry about sticking to the exact letter of the original formulation? But, the motivation for considering cyclic processes is not merely pedantry. The demonic strategy needs to be reliable: if it is going to be more interesting than the 'lucky' demon of Section 2.26.

2.29.3 How reset leads to entropy increase

Discussion of RESET: reset is a logical irreversible operation. That is, it is a many-to-one function. We say that a logical transformation, L, is *logically reversible* if and only if $L : X \to Y$ is a one-to-one (injective) mapping. Hence, with a reversible logical transformation, we can uniquely reconstruct the input state from the output state. If L is not a one-to-one mapping, we say that it is *logically irreversible*. (And this is why Landauer's principle is sometimes discussed in isolation from Maxwell's demon; there is a debate over whether there is a connection between logical and thermodynamic irreversibility. Naturally this debate involves considerations from information theory. One 'cheap connection' that could be made: reset involves throwing away information — and this is the 'throwing away of information' sometimes associated with increasing entropy. But settling whether this is more than superficial connection would involve considering the relationship between information theory and thermodynamics — in particular to give a quantitative analysis to show that not only is there an increase in entropy, but an increase by the right amount — and this is beyond the scope of this thesis.)

Why RESET leads to entropy increase: The memory consists in N DOFs, and these N DOF represent N bits. The Hilbert space for each DOF has two subspaces which represent '0' and '1' (and for simplicity are assumed to be of the same volume). A binary measurement (of the form: is the molecule on the righthand side of the box?) takes the memory DOF which is initially in the 0 subspace to 1 if YES, by implementing $\hat{T}$. But it remains in the '0' if NO, by implementing the identity operator, $\hat{I}$.

$$\hat{V} = \hat{P} \otimes \hat{T} + (\hat{1} - \hat{P}) \otimes \hat{1} \tag{2.59}$$

where $\hat{P}$ is a projective measurement on the system ('is it on the rightside?'). Then the controlled operation on the system can be represented by the following unitary:

$$\hat{U} = \hat{U_0} \otimes \hat{P_0} + \hat{U_1} \otimes \hat{P_1} \tag{2.60}$$

where P_0 is a projection on the 0 subspace — so if the memory is not in the 0 subspace this branch of the unitary gets killed off, thus just implementing the other operation on the system.

Now to consider RESET of the memory. The entropy associated to a memory register with N DOF (or n systems), where the 0 subspace has volume V and the 1 subspace is also V: is $Nln2V$, because the distribution is uniform over all the available space. [53]

Before the reset operation, the entropy of the N bit is $Nln2V$ — and afterwards it is $NlnV$, so there has been an entropy decrease of $ln2$ per bit. But the assumption was that the total evolution of the demon and all the systems were unitary — and thus, phase space preserving. The memory device's entropy decreases and so the environment (the demon computer and all other systems) must have an associated increase in entropy. But "if the computer is to carry out the reset operation repeatably, its own entropy cannot increase without limit. So a repeatable reset process dumps at least entropy $ln2$ per bit into the environment. In the special case where the environment is a heat bath at temperature T Landauer's principle becomes the requirement that reset generates $Tln2$ per bit" (Wallace, 2014, p. 721).

To sum up: Whether something is physically possible, depends on which physical theory we look at. According to thermodynamics, a Maxwell demon is not possible. According CM/QM, it looks like a Maxwell demon is possible. But we need to consider how the Demon operates, because a crucial part of the Second Law is considering how

[53] But one might object: if the memory is *either* in 0 or 1 — which has the same phase space volume as after reset to 0. i.e. in both cases the accessible phase space is V. But if we know the state of the memory device, i.e. if it 0 or 1, then the reset can be implemented without entropy cost, because this is reset of *known* data (cf. (Feynman, 1996, p. 144): "if we know the atom's position, we expend no energy in resetting, irrespective of where the atom starts out". (See (Ladyman and Robertson, 2013, p. 267) for more details). But this would require the state of the memory device to recorded elsewhere in the total system — and this would need to be reset — and so the problem gets pushed back a stage. So we can't know whether the memory device in 0 or 1.

the environment is affected by a given thermodynamic process. Insofar as the total system (i.e. the system including all control mechanisms) is a statistical mechanical system, we know that the Gibbs entropy is non-decreasing. Thus LP explains where the compensating entropy increase comes from.

2.29.4 The Status of Landauer's principle

Earman and Norton (1998, 1999) have sharply criticised the literature on 'saving the SL' with LP.[54] Much of their critique I agree with. For example, I completely agree that arguments aiming to show that thermal fluctuations will frustrate the mechanism of a particular Maxwell demon, still leave it open that there could be another realisation of the demon. But here I dissent from their rejection of LP.

Earman and Norton point that proofs of LP either assume TDSL or do not. If they do, these are 'sound' and merely of pedagogic value. That is, these proofs have shown how TD is internally consistent, but haven't established that the TDSL is not violated sometimes — this has just been assumed from the outset. Alternatively, a proof may not assume SL —– and thus has a chance of 'saving it'. In this case, they call the proof 'profound'. But such a proof would need a new physical principle — which they claim is not forthcoming.

But there is an ambiguity in the phase 'a new physical principle'; does 'new' mean 'never before' or just 'external to TD'? I think the latter reading suffices, since the Maxwell's demon was conjured up to probe a theory we already have: TD. No one thinks Maxwell's demon is a battle cry for a new fundamental physics, but rather it is a device for illuminating our current theories. Thus, whether an argument for LP is sound or profound depends on whether the resources used are external to TD or not.

I think this explains Wallace (2014)'s claim that LP is (a) Sound with respect to SM, yet (b) profound with respect to TD. (a) In Section 2.29.3, we just assumed that the controlling system + controlled system, i.e. the demon (including its memory), the system it acts upon and the surrounding environment, all evolved unitarily. This had the implication that overall Gibbs entropy won't decrease. (b) But this dynamical assumption is not identical to the TDSL. (In Section 2.31.2, I will argue that the TDSL should not be identified with 'SM is not decreasing'. The main idea is: these unitary dynamics also imply that the entropy non-*in*creasing so this dynamical assumption doesn't play the same role as the TDSL). Thus, we have not assumed the TDSL straight up, instead we have used resources from outside of TD, and so in this sense LP is profound with respect to TD.

[54] Feyerabend's, and Popper's critique is similar in spirit to Earman and Norton: all involve charges of circularity. "To sum up: the attempt to save the second law from systematic deviations, apart from being circular and based upon an ambiguous use of the term 'information', is also ill-conceived, for such deviations are in principle possible" Feyerabend (1966).

Thus, we should think of LP as explaining why we can't *reliably* make the TD entropy decrease, or construct an engine more efficient than a Carnot engine— given the assumption that phase space volume can't decrease. This is enlightening because it is not immediate what the connection between SM and TD entropy is (and will be discussed throughout this Chapter). There is a danger that LP is taken to say: if we assume entropy (SM) is non-decreasing, then entropy (TD) is non-decreasing. Of course, this is unenlightening until we see that 'entropy' in the antecedent refers to the SM entropy, and in the consequent 'entropy' refers to the TD entropy. Thus, LP is informative, and not superfluous.

2.30 Conclusion about the scope of the TDSL: the verdict on options 1-4.

To conclude, I have rejected option 4. *(the Maxwellian view)*. We saw that Landauer's principle explains why—despite our improved control of individual molecules since Maxwell's day— we still can't build an engine more efficient than a Carnot engine to solve the world's energy crisis.

We saw that options 1., 2., and 3. come together in the TD limit. Given that macroscopic systems have vastly many constituents ($\sim 10^{23}$), we can see why the TDSL captures something true about our world, despite the initial doubts cast by the atomic nature of matter in Section 2.24. But does this have the implication that the TDSL only applies to macroscopic systems? No, because a demon cannot harness the fluctuations — no matter how small the system is. Instead, it is an open question whether TD applies to small n systems: ultimately, this will depend on whether TD is useful in these domains. In the case of Brownian motion it is unclear whether the concepts of heat and work are applicable to individual particles. But much recent work, e.g. that of Linden et al. (2010) on the smallest possible thermal machines, suggests that TD is useful for small systems.

If we are not compelled to restrict the TDSL to large systems, then we do not have to endorse 3. *(the Statistical view)*. So, what should we conclude about the scope of the TDSL? My verdict respects the motivation for option 2. *(the Probabilistic view)* but emphasises how close this is to the strict view.

Probabilistic fluctuations allow spontaneous differences in temperature, pressure and other macrovariables. But there is no way to harness these fluctuations without feedback processes, or controlled operations, to make an engine more efficient than a Carnot engine. Even with a small system, we cannot create an engine with an efficiency greater than a Carnot engine: because to do so would require feedback or control processes. And LP tells us that such processes have an entropic cost. Thus, given weak assumptions at the SM level (roughly the preservation of phase space volume) we see why despite our ability to manipulate individual molecules we cannot construct an engine more efficient than a Carnot engine.

Nonetheless we still need the 'reliability' caveat: we cannot have the strict TDSL unaltered — there is one counterexample. Earlier I suggested that there could be an ignorant yet 'lucky demon' that doesn't perform a controlled operation, but just guesses which side to insert the piston. Whilst this will not be a reliable method, we have still had one cycle more efficient than Carnot. Indeed, this is the motivation for the 'probabilistic' version: we are very unlikely to get lucky.

Hence, my (uncontroversial) verdict is: the TDSL must be altered to say that we cannot *reliably* transfer heat from cold to hot, or turn heat into work. But perhaps this is not really an *alteration* of the TDSL. This is because the *cyclic* component of the TDSL already captures this: one way to ensure that a process can be implemented reliably is if the final state is the same as the initial state (i.e. a cycle) — because the process can then just be repeated (cf. Wallace (2014)).

2.31 The Image of the TDSL in SM

What is the 'image' of TDSL in SM? Given the difficulty, or unnaturalness, of the distinction between heat and work (discussed in Part I on the First Law), the Kelvin formulation does not have an obvious correlate. Indeed, cyclic processes and Carnot engines do not have the starring role in SM that they do in TD. Transferring heat between bodies of different temperatures is also not the main concern of SM either. Instead (as discussed extensively throughout this thesis), non-equilibrium SM is concerned with qualitatively describing the approach to equilibrium (from non-equilibrium states). And equilibrium SM calculates various macroscopic quantities from an equilibrium probability distribution, such as the canonical ensemble (and the partition function Z plays a starring role). This is the sense in which SM and TD have different subject matters.

Thus, the natural way to connect these two subject matters in order to find the image of the TDSL within SM is this: the TDSL has the implication that the TD entropy cannot decrease (for an isolated system). Hence, finding the realiser of the TD entropy in SM seems key to finding the 'image' of the TDSL in SM. Indeed, Callender (1999) calls this the search for the 'The Holy Grail': find a SM function to call 'entropy' and establish that it is non-decreasing.

Traditionally, the search for the Holy Grail has led to strife — there are (at least) two SM entropies vying for the position: the Boltzmann entropy and the Gibbs entropy. In the next Section I briefly outline why this is so, before rejecting the traditional Holy Grail.

I argue that these considerations are besides the point for the TDSL because I claim that 'non-decreasing SM entropy' is neither necessary nor sufficient for a criterion of the SM realiser. Instead I offer a different criterion which emphasises the importance of quasi-static processes in TD, and will happily fit my overall endorsement of a Gibbsian view.

2.31.1 The Strife and the Holy Grail

The search for the Holy Grail — a non-decreasing SM function to call the 'entropy' — has led to strife because neither the Gibbs nor the Boltzmann entropy match the TD concept exactly. The Boltzmann entropy, S_B, can decrease (as we saw in Section 2.26). The Gibbs entropy, S_G, differs from the TD entropy because it is not a property of an individual system, but a property of the ensemble.[55]

Debate then ensues about which SM entropy is the 'right' microphysical entropy. Callender (2001) argues that the virtues of the Boltzmann entropy outweigh its vices; the fact S_B can decrease unlike S_{TD} should not militate against choosing S_B.

But it seems to me that the traditional strife is misguided for three reasons. (1): It is not clear that Gibbs and Boltzmann entropies must be considered as rivals on my view of inter-theoretic relations. If these entropies are truly distinct (cf. the appendix) contra the discussion in Section 2.6, then they need not be rivals — but different realisations in different theories.

(2): Differences between the TD and SM concepts of entropy should be expected. As I discussed extensively in Chapter 1, two theories will inevitably employ different concepts. They are different theories, after all. Furthermore, in order to secure a reduction, the lower-level theories' quantities must only capture the relevant, or crucial, features of the higher-level theories quantities. I spelt this out as the SM realiser fulfilling the TD role. (The idea being that some features of the TD concepts will not be crucial, and so not part of the role). Thus, if being a categorical, rather than probabilistic property is not a crucial part of the TD entropy role, then this problem can be dismissed.

(3): Of course, this raises the question: which differences are important, i.e. prevent a SM entropy from playing the TD role? Boltzmannians might argue that being a property of the individual system is an essential part of the TD role. But I think that the TD role of S_{TD} as defined by the Holy grail does not capture the right features of the TDSL: I now argue that a non-decreasing SM function is neither necessary or sufficient condition for playing the S_{TD} role.

[55] But, as discussed earlier, this difference arguably disappears in the quantum case; the distinction between the ensemble and the individual system is blurred when the density matrix is taken to be the fundamental description of the system.

2.31.2 Non-decreasing SM entropy is neither necessary nor sufficient

Why should we think that a non-decreasing SM entropy function is the SM realiser of the TDSL? 'SM entropy is non-decreasing' isn't a necessary condition for an SM realiser: sometimes the SM entropy can decrease, but additional assumptions are required for this to raise any issues for TDSL.

The 'thermodynamic entropy law' is a corollary of the TDSL: if TDSL, then S_{TD} is non-decreasing. Thus, if the S_{TD} decreases, then the TDSL is violated. (If NOT-S_{TD} non-decreasing, then NOT-TDSL). But unless the S_{TD} is *identified* with the S_{SM}, it is unclear what the consequences of a decreasing S_{SM} function are for the TDSL. The thermodynamic entropy S_{TD} is only defined at equilibrium: as many have emphasised, it is silent about what happens away from equilibrium. [56] Thus, if a candidate SM entropy decreases sometimes, this needn't be problematic. Hence, 'non-decreasing SM entropy' is not the right desiderata for microphysical realiser of the TDSL. Instead, the SM entropy must not decrease in certain situations: a decreasing SM function only conflicts with the TDSL if the SM entropy decreases between isolated equilibrium states. In particular, we saw in Section 2.23 that the S_{TD} is non-decreasing in quasi-static adiabatic processes — thus, this is the behaviour that an SM realiser must capture. This is the *true* Holy grail.

Whilst I am advocating a Gibbsian view of SM, the fact the S_G is non-decreasing does not suffice. The dynamical fact that S_G is non-decreasing was a key assumption in LP, which makes it tempting to claim that S_G plays the role of S_{TD}. But whilst the fine-grained Gibbs entropy is non-decreasing — but it is also non-increasing, and thus doesn't display the same behaviour as the thermodynamic entropy, S_{TD}. Thus, being 'non-decreasing' is not sufficient: I now discuss an essential part of the S_{TD} role – its behaviour in quasi-static and non-quasi-static changes.

2.31.3 Quasi-static changes in SM

In this Section, I explicate the image of the TDSL in SM. As I have emphasised throughout this Part of the Chapter, slow changes to external parameters are required. This can be understood in quantum SM in the Gibbsian framework, but fits less naturally into a Boltzmannian picture.

The TDSL claims that there is an important difference between quasi-static adiabatic changes and non-quasi-static adiabatic changes to a system: the latter is S_{TD} increasing, whereas in the former S_{TD} is constant. How do we see this distinction in SM? If it is to mirror the TD situation, the S_{SM} must be constant when an external parameter is slowly altered, but increase when the external parameter changes quickly. In this Section, I

[56] Indeed Uffink points out that the Kelvin and Clausius formulations don't say that the entropy must be monotonically increasing.

outline how slow (quasi-static) and rapid (non-quasi-static) changes are considered in quantum statistical mechanics.

In quantum mechanics, if the system is isolated:

$$|\psi(t)\rangle = e^{i\int \hat{H}dt}\,|\psi(0)\rangle\,. \tag{2.61}$$

Alternatively we can describe the systems state with the following density matrix:

$$\rho(t_0) = \Sigma_i p_i\,|\psi_i\rangle\,\langle\psi_i|\,. \tag{2.62}$$

In (Gibbsian) SM, thermal equilibrium is represented by the canonical distribution.[57] The canonical distribution represents the probability of the system being in a given energy eigenstate $\phi_j(t_1)$ ($|\psi_i(0)\rangle = |E_i(0)\rangle$). Each state evolves under the time-independent Schrödinger equation (and so is unchanging). As such, the density matrix is unchanging in time too:

$$\hat{\rho} = \Sigma_i \omega_{ii}\,|E_i(0)\rangle\,\langle E_i(0)| \tag{2.63}$$

But once we change an external parameter at $t > t_1$ (such as the volume of the box), we have a time-dependent energy operator. Thus, the state of the system will be changing over time: $|\psi_i(t)\rangle \neq |E_i(0)\rangle$. In what I follows I first consider what happens when the external parameter is changed quasi-statically, and then when it is changed rapidly.

Slow change: As discussed extensively earlier, a quasi-static process is one where the external parameters are changed so slowly that the system is approximately in equilibrium. Taking this expectation from TD, the hope is that if the external parameter is changed slowly enough, the system will remain approximately in the canonical ensemble, i.e. 'close to equilibrium'. And this expectation is correct, provided we can establish:

1. $\phi_j(t)$ is an energy eigenstate of $H(t)$ for $t > t_1$, i.e. $|\psi_i(t)\rangle = |E_i(t)\rangle$
2. The probability distribution has the form of the canonical distribution: i.e. an exponential dependence on the energy eigenvalue.

1. is established by Ehrenfest's principle:

[57] The microcanonical and grand canonical distribution can also represent thermal equilibria. As discussed in Chapter 4, in principle, the choice of distribution depends on certain features of the situation: can the system exchange energy with the environment? Can the particle number change? But for all practical purposes, in large N situations, the distributions can be used interchangeably, and the canonical distribution is often the workhorse of equilibrium SM.

> Ehrenfest's Principle: If the energy eigenstates of $H(t)$ are non degenerate for times $t > t_1$, if $\phi_j(t_1)$ is an energy eigenstate of $H(t_1)$, if $\phi_j(t)$ is the state evolved from $\phi_j(t_1)$ according to the Schrödinger equation, and if the external parameter changes very slowly, then $\phi_j(t)$, for each time $t > t_1$, is very nearly an energy eigenstate of $H(t)$ at the corresponding time. In the mathematical limit of a finite change in the external parameter occurring over an infinite time interval, "is very nearly" becomes "is" (Baierlein, 1971, p. 380).

For this 'slow change' in an external parameter, $\hat{\rho} = \Sigma_i \omega_{ii} \left|E_i(t)\right\rangle \left\langle E_i(t)\right|$ (where elsewhere $\omega_{ii} = p_i$). This means that the energy eigenstates change over time, but the changes to the Hamiltonian do not take earlier energy eigenstates to later non-eigenstates.

2. Whilst there is no general proof of 2., Baierlein (1971) motivates why it is a reasonable assumption, as follows.

Each state $\phi_j(t)$ alters in the interval $t > t_1$, but the probability p_j assigned to it stays the same: in this way, a given eigenstate carries its probability with it. As the energy eigenstates shift, the distribution shifts (as seen on the right hand of the figure 2.8). The question is: does this new distribution have the canonical form? For the distribution to re-arrange into a gaussian distribution, the eigenstates would need to cross (i.e. the originally high probability lowest energy eigenstate must be shifted to a much higher energy eigenstate, in order to be the peak of the gaussian distribution). But as Baierlein says, if there is no degeneracy no vertical lines will cross each other, and so such radically different distributions are not possible.

But we might think that even if such radical changes to the distribution are not possible, why think the new distribution is 'suitably close' to the canonical distribution? Whilst a hard and fast proof is not available, there are two heuristic reasons to think it will be: 1. temperature at the later time is chosen using constancy of S_G, and thus exponential curve placed at the right height. 2. Even if the distribution differs from the canonical distribution for some energies ('out in the tails'), only the states near $\langle E\rangle$ are important for estimating the macro properties, so provided that the distribution matches the canonical distribution near $\langle E\rangle$, this approximation will be successful.

Rapid change: when an external parameter changes rapidly, the system does *not* remain close to equilibrium, and there is no reason to expect a probability distribution of canonical form to apply during the change of the external parameter. In this case, $\rho(t) = p_i \left|\psi_i(t)\right\rangle \left\langle \psi_i(t)\right|$, where $\left|\psi_i(t)\right\rangle \neq \left|E_i(t)\right\rangle$. If we were to write $\rho(t)$ in the energy eigenbasis, we would see that the density matrix is not diagonal in this basis: $\rho(t) = \Sigma_{ij}\omega_{ij} \left|E_i(t)\right\rangle \left\langle E_j(t)\right|$.

But when the external parameter is no longer varying, we expect the system to reach

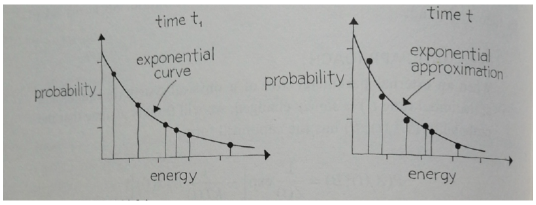

Figure 2.8: The exponential approximation. Diagram from (Baierlein, 1971, p. 385).

a new equilibrium, i.e. canonical distribution. Of course, justifying this is controversial, and is part of finding the underpinning to the Minus First Law. Discussing the topic of the approach the equilibrium (and the associated entropy S_G increase) is the project of Chapter 3. For now, we follow the pragmatic move of Baierlein, and leave the justification of this practice until the next Chapter.

The pragmatic move is just to *adopt* a new canonical distribution with energy eigenstates appropriate for the new volume. In other words, we coarse-grain:

$$\rho = \Sigma_{ij}\omega_{ij} \left|E_i(t)\right\rangle \left\langle E_j(t)\right| \to \rho_{cg} = \Sigma_i \omega_{ii} \left|E_i(t)\right\rangle \left\langle E_i(t)\right|, \tag{2.64}$$

where we assume that the off-diagonal terms $w_{ij}, i \neq j$ are small so $\Sigma_{ij}\omega_{ij} \left|E_i(t)\right\rangle \left\langle E_j(t)\right| \approx \Sigma_i \omega_{ii} \left|E_i(t)\right\rangle \left\langle E_i(t)\right|$, where t is a long time after the external parameter has stopped changing.

Change in Entropy: In the slow change, the probability distribution evolves according to the microdynamics, and so $S_G = -k_B \int \rho ln\rho$ is constant. In the rapid change, we just adopt a new probability distribution, ρ_{cg}, which is distinct from the evolved distribution, ρ: this adopted distribution has discarded all the information about the initial conditions and correlations, and so S_G has increased.

Whilst it is tempting to claim that the original distribution ρ is the *true* distribution, and so S_G has not really increased, in the next Chapter I will argue that this is not the case: ρ_{cg} is not a 'distorted' or false distribution.

Thus, we can find the image of the TDSL in QSM, by considering the canonical ensemble and the quantum adiabatic theorem, but we still need to rely on certain controversial issues in non-equilibrium SM (i.e. that during the approach to equilibrium the Gibbs entropy increases).

The above discussion employs a Gibbsian picture of SM. I now attempt to connect this discussion to the Boltzmannian framework, but conclude it is less natural. In the Boltzmannian picture, the measure is defined over the available energy hypersurface. In a TD process, such as a quasi-static adiabatic expansion, the available energy hypersurface changes — and so the area of which the measure is defined changes. Insofar as the Boltzmannian measure *is* just one of the Gibbsian ensembles, perhaps the above account can be carried across. But the above account drew a crucial distinction between just redefining the ensemble (or in Boltzmannian terms: the measure) and the ensemble evolving according to the microdynamics. On a Boltzmannian picture, both sides of the distinction just look as if we are redefining the measure — and so it is hard to see how import this account into the Boltzmannian framework. In this respect, this account tells in favour of a Gibbsian account.

2.32 Conclusion

The project of this part of the thesis has been to see if TD can be reduced to SM: which means — is there an SM realiser of the TD role? The first Sections of this part worked on articulating this role: I emphasised three things (i) the different types of irreversibility in play (ii) the distinction between the Minus First Law and the Second Law and (iii) the importance of the environment.

The discovery of the atomic nature of matter might have lead to the belief that the TDSL is false — because of the demons arise from the depths of the lower-level theory. I considered in what sense a Maxwellian demon is possible. Of course according to TD, such a demon is not possible, whereas the existence of such a demon looks plausible according to CM/QM. Yet Landauer's principle shows that — provided the mechanism of the demon obeys SM assumptions — such a demon is not possible.

Whilst the TDSL is not false, one might have nonetheless thought its scope needed to be limited, as a result of fluctuations. I argued that the considerations about the scope of the TDSL and whether it is a mere 'statistical truth' were often concerned with the Minus First law — the approach to equilibrium, rather than the TDSL. The fluctuations which throw a shadow of doubt on the TDSL disappear in the thermodynamic limit, suggesting that provided we are talking about macroscopic systems, the TDSL applies. The scope of the TDSL only had to be altered to include a 'reliability' caveat.

Because of the different concepts — most importantly heat and work — in thermodynamics, finding a correlate to the classic formulations of the Second Law is not straightforward. The natural tendency is to try to find a non-decreasing entropy function: what Callender dubs the search for the Holy Grail. In the literature, the concern is then whether the Boltzmann or the Gibbs entropy is the right candidate. But I argued that this is neither necessary nor sufficient: instead we need an entropy function that is constant during quasi-static adiabatic processes and increasing in non-quasi-static processes.

In order to find a SM realiser of this TD entropy role, I took a Gibbsian approach. By using Ehrenfest's adiabatic principle, we found that the Gibbs entropy is constant during a quasi-static adiabatic process, but increased during a non-quasi static, i.e. rapid, process.

Part V. The Third Law

There are many formulations of the Third Law. Here I will not discuss the unattainability principle: the claim that it is impossible to lower the temperature of a system to $T = 0$ in a finite number of steps (Masanes and Oppenheim, 2017). Instead, I focus on the 'heat theorem' also known as The Nernst Postulate. According to this formulation, the Third Law states that:

$$lim_{T \to 0} S(T, X) = 0 \tag{2.65}$$

where X represents the variables, other than T, that S depends on. This law says that as the temperature tends to absolute zero, the entropy tends to zero. Often this statement is weakened: the entropy need not be zero, just some finite constant. For example, the requirement becomes that the entropy density S/N vanishes as $T \to 0$ and $N \to \infty$ (and so a finite S is permitted). Thus: $S/N \to 0$ as $T \to 0$ and $N \to \infty$.

All hands agree that the Third Law, also known as Nernst's postulate, has less bite than the other laws — unlike those, it does not implicitly define a new quantity. Instead, the Third Law defines an absolute scale for entropy, as the Second Law only defined entropy *differences*. (Recall from Part IV that the entropy was only defined relationally, with respect to a given reference state.) Why think that defining an absolute scale for entropy is an important thing to do? This question is especially pressing given that there is much debate over the physical content of the Third Law — and whether it is violated. For example, Wald (1997) disputes whether the Nernst postulate should even be considered to be a real law.[58]

Defining an absolute scale for entropy, which is what is achieved if the TD entropy is zero at $T = 0$, means that entropy is no longer a relational quantity as it is now absolute.

Why is this significant? It means that we can compare entropies across disparate systems. As we saw in Part IV, there has to be a quasi-static process connecting two states in order to define an entropy difference between them. Thus, the entropy difference is only defined between states in the same state-space, Ξ.[59] Because of this, it is hard to see how to compare the TD entropy of a system such as a gas with the TD entropy of a magnet — their respective states live in different spaces and so there is no common reference state. The significance of this point is opaque to me: do we ever need to compare the TD entropies of a gas and a magnet? Practically speaking, perhaps not. But if we cannot compare across the entropy of disparate systems, this tells against TD entropy as being a universally applicable property.

A consequence of the Third Law is that the heat capacities must tend to zero at $T = 0$.

[58] Wald (1997, p. 1) says that "the main of this paper is to attempt to lay to rest the "Nernst theorem" as a law of thermodynamics."

[59] I am grateful to Erik Curiel for this point.

This is because:[60]

$$S(B) - S(A) = \int_A^B dT \frac{C_v}{T} \tag{2.66}$$

which comes from

$$\frac{\partial S}{\partial T} = \frac{\partial S}{\partial E} . \frac{\partial E}{\partial T} = \frac{C_v}{T} . \tag{2.67}$$

If the entropy at $T = 0$ is finite, then the lefthand side of equation (2.66) must be finite — implying that the righthand side must be too: the integral must converge, and so C_v must head to zero at least as quickly as T does. Thus, as $T \to 0$, $C_v \to T^n$ for $n \geq 1$.

We can examine individual cases to check whether this is true. First let us take the ideal gas. Since $E = \frac{3}{2}k_B T$, $C_v = \frac{3}{2}k_B$. Because this is a constant, it obviously does not tend to zero. But this is unsurprising: we already saw that the ideal gas model breaks down at low temperatures (as seen in Part II). Yet the low temperature problems are not limited to the ideal gas case. Other classical models, such as the Dulong-Petitt model have a constant heat capacity, and so face the same problem.

But the condition $T \to 0$, $C_v \to T^n$ for $n \geq 1$ does hold in quantum models. For example, the Deybe model of solids models the phonon contribution to the heat capacity in such a way that $C_v \propto T^3$. The electrons in a metal can be treated as a Fermi gas, for which $C_v \propto T$. Thus, Tong (2012, p. 134) claims that "the Third Law is an admission that the low temperature world is not classical. It is quantum." That is: there is a QSM, but not CSM realiser of the Third Law.

Of course, now the question is whether the Third Law always holds for the quantum, beyond the examples I have cited. Wald claims that whilst the Third Law holds for many quantum systems, there are nonetheless counterexamples. But the QSM realiser of the Third Law need not be exceptionless. Nonetheless, Tong (2012) gives some generic reasons for thinking that QSM can explain the Third Law, as follows.

If the ground state is degenerate, there will be many microstates corresponding to the same energy eigenvalue — and so S is non-zero. (If there were only one microstate corresponding to the ground state, where $E = E_0$, then the entropy is zero: $S = -k_B log \Omega = -k_B log 1 = 0$). For a degenerate ground state, there is more than one energy eigenstate corresponding to the lowest energy eigenvalue E_0. But is worrying? After all, we considered the weakened Third law, where $S \to k$ as $T \to 0$. Does a degenerate ground state violate this weakened Third Law? It could do. It is not guaranteed that the entropy will tend to a constant as $T \to 0$: the ground state could be *more* degenerate than the other energy eigenstates. And so rather than $S \to k$, the entropy could increase as $E \to E_0$ tends to the ground state.

But Tong (2012, p. 134) claims there are generic mathematical reasons to think that the ground state will not be degenerate for a *large* system. It is hard for large matrices

[60] Equation 2.66 provides the link between theory and experiment — if we can measure the heat capacity at different temperatures then we can work out the entropy changes.

to be degenerate, as any nonzero off-diagonal term — no matter how small — will lift the degeneracy. Thus, if the ground state is non-degenerate, there will be only one microstate associated to it, which ensures that $S \to 0$, as $T \to 0$.

2.33 Conclusion: TD reduces to SM

I claimed that TD would be reduced to SM if a SM quantity could be found that plays the nomological role of the quantities implicitly defined by the laws of TD: (i) temperature, (ii) energy and (iii) entropy.

(i) For the case of temperature, I claimed that $\frac{\partial E}{\partial S}$ plays the role of temperature: it is the property which two systems have in common if they are in equilibrium with one another.

(ii) The case of energy looked deceptively easy, since the conservation of energy is a basic feature of the microdynamics underlying SM. But I argued that the SM realiser must also capture the distinction between heat and work. Within QSM we found the realiser of the First Law: $\Delta E_s = \Delta W_s + \Delta Q$, where $\Delta W = \int_0^\tau \left\langle \frac{\partial H}{\partial t} \right\rangle_\rho dt$, and $\Delta Q = \int_0^\tau \langle [H_s, V_{SE}] \rangle_\rho dt$ where V_{SE} represents the interaction between the system and the environment.

(iii) Finding the SM realiser of the thermodynamic entropy required that first we deflate some of the hype surrounding the TDSL. In particular, I emphasised the importance of quasi-static processes in TD: 'slow' or 'gentle' interventions, rather than uncontrolled interventions on the system. I claimed that the essential role of S_{TD} is that it is constant during a quasi-static adiabatic process but increases during a non-quasi-static process. I argued that the Gibbs entropy $S_G = \int -k_B \rho ln \rho$ plays this role.

One general theme of this Chapter is that it is often perspicuous to consider *quantum* rather than classical SM — breaking with tradition in much of the philosophy of physics literature, which often only considers the classical case. Arguably, this is because much of the literature focuses on the approach to equilibrium (considered in the next Chapter). There, the problem (and its solution) of the approach to equilibrium takes the same form in both QSM and CSM.

But TD describes what happens once the system is *in* thermal equilibrium. Considering external interventions on a thermally isolated system is naturally understood in QSM as intervening on the external parameters, such as volume, in the Hamiltonian which then defines the available energy eigenstates. But in classical case, there is no potential term in the Hamiltonian of the ideal gas, so it is less clear how to think about such interventions. The ideal gas is — as the name suggests — an idealised system, but one that can be derived from the quantum partition function under various assumptions. Considering this derivation allowed us to see when this idealisation broke down: when the gas is cold or dense.

A second, related, general theme: SM and TD have —to some extent— different subject matters. The above example of quasi-static processes is a case in point: neither equilibrium nor non-equilibrium SM focuses on quasi-static processes, and so connecting these different subject matters required some work: merely finding a non-decreasing entropy function would not suffice.

Finally, insofar as we can say that there is a 'received view' of the reduction of TD to SM, one might caricature it as follows: TD reduces to SM through bridge laws such as $T =< K >$. Yet I have claimed that we can find the realisers of the functional roles of TD quantities in SM and thus: TD reduces to SM — in my sense of the word — but not in the way commonly assumed.

2.34 Appendix: Gibbs and Boltzmann entropy

The Boltzmann and Gibbs frameworks are undeniably conceptually different. There are (at least) three ways in which they differ: (i) their definitions of equilibrium, (ii) their definitions of entropy and (iii) their object of study (Lavis, 2005). Nonetheless, their entropies are numerically identical at their respective equilibria. Furthermore, the two entropies are inter-derivable, as I now show. The philosophical significance of their inter-derivability explains why there are no ramifications for the physicists' lackadaisical attitude to these conceptual differences.

The Boltzmann entropy measures the number of microstates compatible with the current macrostate the system is in:

$$S_B(E) = k_B ln\Omega(E), \tag{2.68}$$

where $\Omega(E)$ is the number of microstates of the system with energy E. This is often motivated by considering the combinatorics of, e.g. coin flipping: there are 45 possible ways to get 2 heads in 10 coin flips. Likewise, there are various arrangements of molecules. But often the possible states are not discrete like coin flip case, and so a measure μ is defined on the phase space. The Boltzmannian definition of equilibrium is the largest macrostate [61].

In contrast, the starting point for Gibbs entropy is a probability distribution. Hence, the object of study is claimed not to be the individual system (as in the case of Boltzmann) but the 'ensemble', i.e. the probability distribution.[62] This probability will be stationary at equilibrium: $\frac{\partial p}{\partial t} = 0$.

[61] However, see Werndl and Frigg (2015a,b) for an alternative definition.

[62] Again, this distinction between the state of the individual system, and the probability distribution is deflated if ρ is taken to be the density matrix and to represent an individual system.

$$S_G(p_i) = -k_B \Sigma_n p_i lnp_i \tag{2.69}$$

In textbook presentations, the two frameworks are blurred together by the so-called *fundamental assumption of statistical mechanics*: each microstate is equally likely. The probability that the system with fixed energy E is in a given state $|n\rangle$ is then simply:

$$p(n) = \frac{1}{\Omega(E)} \tag{2.70}$$

This is the microcanonical distribution: i.e. the uniform distribution over the available energy hypersurface.

From Gibbs to Boltzmann: The Boltzmann entropy can be recovered from the Gibbs formula:

$$S(p) = -k_B \Sigma^n p(n) lnp(n) \tag{2.71}$$

Using equation 2.70,

$$S(p) = -k_B \Sigma_n \frac{1}{\Omega} ln \frac{1}{\Omega} = k_B \frac{n}{\Omega} ln\Omega \tag{2.72}$$

n labels the number of possible states $|n\rangle$ and we know the number of states is Ω: thus we recover equation 2.68.

From Boltzmann to Gibbs: We can also go from the Boltzmann entropy to the Gibbs entropy (without the assumption that each state is equally likely, i.e. that the probability distribution to be used is the microcanonical distribution).

Assume we have W identical copies of the states, where W is very large.[63] We then can say that the number of systems in state $|n\rangle$ is $p(n)W$. How many ways are there to put $p(n).W$ systems into state $|n\rangle$ for each n? This gives us the number of microstates.

$$\Omega = \frac{W!}{\prod_n (p(n)W)!} \tag{2.73}$$

We then use Stirling's formula:

$$lnN! = NlnN - N + \frac{1}{2} ln 2\pi N + \mathcal{O}(\frac{1}{N}) \tag{2.74}$$

We will only use the first two terms of this approximation.

$$S = k_B ln \frac{W!}{\prod_n (p(n)W)!} = k_B (WlnW - W - ln \prod_n (p(n)W)!) \tag{2.75}$$

$$S = k_B (WlnW - W - \Sigma_n (p(n)Wln(p(n)W - p(n)W) \tag{2.76}$$

[63] In this case, the 'large N' assumption allows us to go from probabilities to actualities.

$\Sigma_n p(n)W = W$ so the second and fourth terms cancel.

$$S = k_B(WlnW - \Sigma_n(p(n)Wln(p(n)W) = k_B(WlnW - \Sigma_n[p(n)Wlnp(n) + p(n)WlnW)] \tag{2.77}$$

For the same reason, the first and third terms now cancel. Thus we end with an expression for the entropy S of the ensemble of W systems.

$$S = -k_B \Sigma_n p(n)Wlnp(n) \tag{2.78}$$

Because S is extensive, we can now divide by W to get the entropy of the individual system:

$$S = -k_B \Sigma_n p(n)lnp(n). \tag{2.79}$$

This is the familiar Gibbs entropy.[64]

Thus, Boltzmann and Gibbs entropies are, in a sense, interderivable but conceptually distinct.

[64] More generally, when the probability distribution can be representing anything, this is the Shannon information entropy.

3 Asymmetry, abstraction and autonomy: justifying coarse-graining in statistical mechanics

3.1 Introduction

Many processes occur in only one direction of time. People age, buildings crumble, eggs smash and gases spontaneously expand — towards the future. Rewinding a film of such processes displays an unphysical sequence of events: eggs cannot unsmash and people cannot become younger. A more technical way of describing the 'directedness' of such processes is to say that the laws governing these processes are not time-reversal invariant (TRI). That is, the time-reversal operator $\mathcal{T}$ does not send solutions of the equations — i.e. histories of the systems at issue — to solutions. (The time-reversal operator varies across theories, but here I take $\mathcal{T}$ to be the map $t \mapsto -t$.)

In stark contrast, the laws of fundamental physics are TRI.[1] The two sequences of events displayed by a film playing forwards, and in rewind, are both physical possibilities. That is, they are both solutions to the laws of fundamental physics. This leads to a traditional problem: given that the fundamental laws are taken to underpin all other processes, how can the fundamental time-symmetry be reconciled with the asymmetry manifest elsewhere?

It is not only the processes of our everyday experience that are irreversible; many equations within physics are also irreversible.[2] In particular, many equations in statistical physics are irreversible, such as the Boltzmann equation, the Langevin equation, the Pauli master equation...the list goes on.

But within statistical mechanics (SM), much progress has been made with this traditional problem. The irreversible behaviour exhibited in non-equilibrium SM can be described by equations collectively called 'master equations', which give 'a purposefully incomplete account of the conservative evolution of some underlying microscopic

[1] Well almost: the relevant symmetry is the CPT-invariance. But the failure of TR-invariance in subatomic physics doesn't underpin the asymmetries discussed here. For the subtleties of TRI, cf. e.g. Roberts (2013, 2017).

[2] Throughout this chapter I take 'irreversible' to mean non-TRI: in the terminology of chapter 2, *irreversibility*$_T$ rather *irreversibility*$_Q$ or *irreversibility*$_R$.

systems' (Liu and Emch, 2002, p. 479). This chapter focuses on one framework, originating in the work of Zwanzig (1960). The idea is that the irreversible equations of SM can be *constructed* from the reversible equations (of either classical or quantum mechanics). I will dub this the 'Zwanzig-Zeh-Wallace' (ZZW) framework, since Zeh and Wallace are prominent later authors who have developed this framework.

However, this framework depends upon the procedure of coarse-graining, which has been heavily criticised. Redhead describes coarse-graining as "one of the most deceitful artifices I have ever come across in theoretical physics" (Redhead, 1996, p. 31) as quoted in (Uffink, 2010, p. 197). Amongst the list of accusations against coarse-graining are: protests of empirical inadequacy, subjectivity and incompatibility with scientific realism. So, if this construction method is to solve the puzzle of time-asymmetry in SM, a justification for coarse-graining is needed. The project of this chapter is to give such a justification.

3.1.1 Prospectus

I will answer two objections to coarse-graining in statistical mechanics. In Section 3.2, I expound the ZZW framework and in Section 3.3, I consider why this framework works. Then I discuss two objections to coarse-graining, namely that the asymmetry resulting from coarse-graining is illusory and/or anthropocentric. Section 3.4.1 outlines these two objections in detail. Section 3.4.2 describes the most prevalent —and I argue unsatisfactory— justification of coarse-graining in the literature, the measurement imprecision justification, which lies behind these objections. In Section 3.5, I outline my alternative justification of coarse-graining which can answer the two objections: these answers are given in Section 3.6 and 3.7 respectively. In Section 3.8, I draw some broader consequences from this alternative justification: the coarse-grained asymmetry is weakly emergent.

3.2 The ZZW framework

The ZZW framework provides a recipe for constructing irreversible dynamics from the underlying reversible dynamics. This framework works with both quantum and classical mechanics (Zwanzig, 1961), although I mainly discuss the classical case. It is clearest to see the framework as constructing an irreversible equation in three stages. First: move to the ensemble variant of the underlying microdynamics. Second: pick a coarse-graining projection $\hat{P}$, whose nature will be described below. Third: two moves are required to find an irreversible and autonomous equation for the coarse-grained probability density.

Stage 1: In classical SM, the state of an individual system is represented by a point in a phase space, Γ-space. (For N particles without internal degrees of freedom, Γ-space is $6N$-dimensional). The system's evolution is determined by Hamilton's equations. However, there is also an ensemble variant of this description. Here probability densities over Γ-space, ρ, evolve according to Liouville's equation, which, like Hamilton's equations, is TRI.[3]

Stage 2: The concept of coarse-graining was originally introduced in a specific form by Gibbs (1903) which I first recall, before describing the *generalised coarse-graining projections* used by the ZZW framework.

Gibbs proposes that the accessible phase-space Γ is partitioned into small, finite volume elements ΔV_m. The coarse-grained density $\rho_{cg}(q, p)$ is then defined by averaging the original probability density $\rho(q, p)$ in each of these boxes. So coarse-graining throws away the information about how exactly the ensemble is distributed across each box.

Gibbs describes the evolution of the probability density by analogy with an ink drop. Dropping blue ink into a glass of water results in the whole glass appearing light blue. However, a drop of ink is an incompressible fluid and so its volume is constant. Upon examination under a microscope, we would see the drop of ink has just fibrillated into thin filaments across the whole glass: cf. Figure 1. So Gibbs' idea is that like an incompressible fluid, ρ often fibrillates over the accessible phase space, as it evolves under the Liouvillean dynamics.

But because ρ behaves like an incompressible fluid, its volume is constant despite its fibrillation; and hence its Gibbs fine-grained entropy, $S_{fg} = -k_B \int_\Gamma \rho \ln \rho d^{3N}qd^{3N}p$ where k_B denotes Boltzmann's constant, is constant. Traditionally, this has been considered problematic, as the thermodynamic entropy *increases*. However, in a coarse-grained description, the density spreads smoothly throughout the available space, and this is well modelled by the coarse-grained probability density, ρ_{cg}. This density has a different entropy, the Gibbs coarse-grained entropy,

$$S_{cg} = -k_B \int_\Gamma \rho_{cg} \ln \rho_{cg} d^{3N}qd^{3N}p. \tag{3.1}$$

Unlike its fine-grained counterpart, S_{cg} *can* increase.

Again, the ink analogy illuminates the discussion of time-evolution. From a macroscopic perspective, the ink smoothly spreads throughout the glass. In the SM case, this 'smooth spreading' of the coarse-grained density ρ_{cg} is described by a 'coarse-grained dynamics', defined as follows. ρ_{cg} evolves forward according to the usual Liouvillean dynamics for a small time interval Δt; and then it is coarse-grained; and this two-step

[3] Throughout this chapter I leave the interpretation of such probability densities open; but admittedly, their connection to the behaviour of individual systems is an urgent issue in the philosophy of SM (see e.g. (Sklar, 1993, ch.3)).

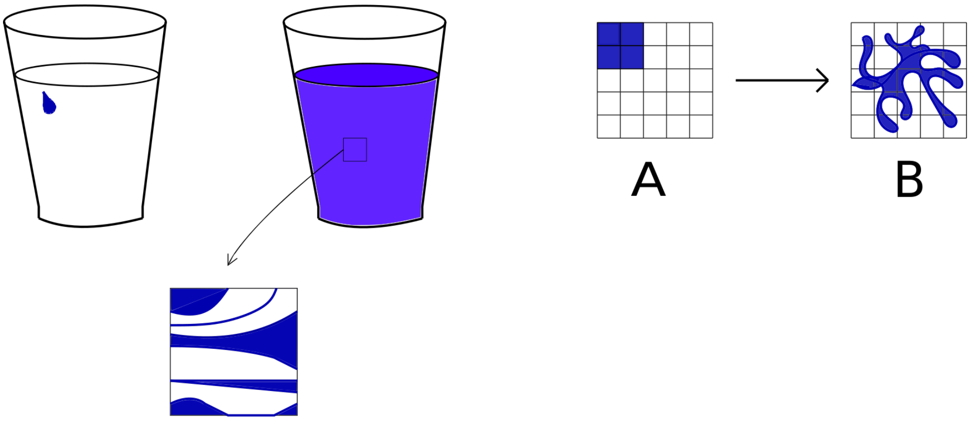

Figure 3.1: A drop of ink in a glass of water fibrillates throughout the whole volume, making the water look blue on a coarse-grained level (pictured on the left hand side). Likewise, a probability density initially concentrated in one corner fibrillates across the available phase space (Sklar, 1993).

process is iterated. This gives what Wallace (2011) terms the *coarse-grained forward (C^+) dynamics* (a label I henceforth adopt).

Note, however, that we could equally well have defined the *coarse-grained backwards (C^-) dynamics* according to which ρ_{cg} is evolved backwards for Δt by the Liouvillean dynamics; and then coarse-grained, then evolved backwards again; and so on. However, this C^- dynamics describes anti-thermodynamic trajectories (where entropy increases into the past) and so is "empirically disastrous". The extent to which the success of the coarse-grained forwards, but not backwards, dynamics can be explained (in particular by appealing to cosmological considerations, such as postulating a 'Past Hypothesis') is controversial (see Earman (2006); Wallace (2011); Albert (2000, Ch. 4)). But in this chapter, it will suffice to admit that the asymmetry has been added in here 'by hand' and thus that this project does not involve locating the 'ultimate source' of the time-asymmetry. For as announced in Section 3.1, I aim only to defend coarse-graining from various objections.

So far, I have only described Gibbs' original coarse-graining. But in the ZZW framework, a more general notion of coarse-graining is used. A coarse-graining projection, $\hat{P}$, acts on the space of possible probability density functions.[4] The important function of $\hat{P}$ is to split ρ into a *relevant part* ρ_r and an *irrelevant part* ρ_{ir}.

$$\hat{P}\rho =: \rho_r, \qquad (1-\hat{P})\rho =: \rho_{ir} \qquad \text{so that} \qquad \rho = \rho_r + \rho_{ir}. \tag{3.2}$$

Here are three examples of a coarse-graining projection $\hat{P}$ defining a relevant density ρ_r. In these examples, the density is defined over a reduced number of degrees of freedom of the systems. Hence we speak of *relevant degrees of freedom*, as well as *relevant*

[4] $\hat{P}$ is idempotent: $\hat{P}^2 = \hat{P}$. $\hat{P}$ is usually linear and time-independent and so commutes with $\frac{\partial}{\partial t}$.

densities.

1. The archetypal Gibbsian coarse-graining discussed above can be written as a projection, $\hat{P}_{cg}$. $\hat{P}_{cg}$ averages over small, finite volume elements ΔV_m $(m = 1, 2...)$ which cover the $6N$-dimensional phase space Γ. These volume elements ΔV_m are sometimes referred to as 'coarse-grained boxes' or 'cells of a partition'. (I write 'ΔV_m' both for the region, and its volume.) Thus for $(q, p) \in \Delta V_m$, i.e. the m^{th} cell, we have

   $$\hat{P}_{cg}\rho(q,p) := \rho_{cg}(q,p) := \frac{1}{\Delta V_m}\int_{\Delta V_m}\rho(q',p')dq'dp' =: \frac{\rho_m}{\Delta V_m}, \tag{3.3}$$

   so that for a general (q, p) we sum over the cells with characteristic functions

   $$\hat{P}_{cg}\rho(q,p) := \rho_{cg}(q,p) := \sum_m \chi_{\Delta V_m}(q,p).\frac{\rho_m}{\Delta V_m} \tag{3.4}$$

   The action of $\hat{P}_{cg}$ is to smooth the density ρ to be uniform across each box, whilst leaving the probability of being in any single box invariant; for all m, $\int_{\Delta V_m}\hat{P}_{cg}\rho = \int_{\Delta V_m}\rho$.

2. Correlations between particles are discarded by appropriate integration, i.e. by taking a marginal distribution. And this can be thought of as applying a projection $\hat{P}_{\mu}$. This projection takes you from a probability density on the full phase space, Γ-space ($6N$-dimensional for N point particles), to the one-particle marginal density, which describes the probability that particle i will be at a particular point in (6-dimensional) μ-space, i.e. have a given $(\vec{q}, \vec{p}) \in \mathbb{R}^6$. Thus, the mapping from Γ-space densities to μ-space densities destroys information about the correlations between different particles and cannot be inverted.

   More generally, in the BBGKY hierarchy we define a system of correlation functions, where f_s gives the probability that s particles have a given position and momenta. Generally the evolution of f_s depends on f_{s+1}, and f_{s+1} depends on f_{s+2}... all the way to f_N (where N is the totally number of particles). But —under certain physical conditions— this chain of equations can be truncated at a given point, i.e. all correlations beyond the three-particle correlations can be thrown away (Huang, 1987, p. 65).

   A projection akin to $\hat{P}_{\mu}$ is used in constructing the Boltzmann equation (see Wallace (2015b, p. 292), (Zeh, 2007, p. 59) and, for an explicit construction of the Prigogine-Brout equation—a cousin of the Boltzmann equation—see (Zwanzig, 1960, p. 1340)).

3. The diagonalisation projection $\hat{P}_{dia}$ applies to quantum systems and removes off-diagonal elements of the density matrix (with respect to some chosen basis). This partitioning into diagonal and off-diagonal matrix-elements (relevant and irrelevant respectively) is used in the derivation of the Pauli master equation (Zwanzig, 1960, p. 1339), where discarding the off-diagonal elements amounts to ignoring interference terms.

Given a coarse-graining projection $\hat{P}$, the next aim is to find an equation for just the relevant degrees of freedom described by ρ_r. By re-arranging the Liouville equation in terms of the two densities, ρ_r and ρ_{ir}, we find the pre-master equation (see (Zwanzig, 1960, §2) for the steps to the pre-master equation);

$$\frac{\partial \rho_r(t)}{\partial t} = \hat{F}\rho_{ir}(t_0) + \int_{t_0}^{t} dt' \hat{G}(t')\rho_r(t-t'), \tag{3.5}$$

where $\hat{F} := \hat{P}Le^{-it(1-\hat{P})L}$ and $\hat{G}(t') := \hat{P}Le^{it'(1-\hat{P})L}(1-\hat{P})L$. L represents the Liouvillean evolution.

This *pre-master* equation is formally exact and so the time-reversibility remains. The first term on the RHS depends on the irrelevant degrees of freedom, ρ_{ir}. The second term is *non-Markovian*; the evolution of ρ_r at t depends on the history of the system between t_0 and t as evidenced by the integral between $t' = t_0$ and t. This is unlike classical mechanical trajectories for which, given the current state, the future evolution is determined without any information about the system's history.

Stage 3: Next, two assumptions are used to arrive at an autonomous and irreversible equation for the relevant degrees of freedom. 'Autonomy' requires that the dynamical evolution of ρ_r has no explicit dependence on ρ_{ir} or t.[5] The reversible pre-master equation (3.5) is of the form $\frac{\partial \rho_r(t)}{\partial t} = f(\rho_r(t), \rho_{ir}(t), t)$ and so is not a time-independent or autonomous equation.

In general, an autonomous dynamics for ρ_r is in no way guaranteed; since ρ can be decomposed any way we like, the aspects of ρ we have dubbed 'relevant' (ρ_r) need not be dynamically autonomous or independent from the irrelevant aspects. Two steps are required:

1. **The initial state assumption** states that the first term vanishes. This is achieved by stipulating that $\rho_{ir}(t_0) = 0$.[6] When $\rho_{ir}(t_0) = 0$, equation (3.5) becomes a closed equation for $\rho_r(t)$.

[5] The condition for an equation to be autonomous, familiar from mathematics, is that "t does not occur explicitly in the equation, as in $\frac{dy}{dt} = f(y)$" (Robinson, 2004, p. 13). This is required so that $\frac{\partial \rho_r(t)}{\partial t}$ has no 'covert dependence' on ρ_{ir}.

[6] This is a sufficient but not necessary condition for this term to vanish; the action of $\hat{P}Le^{-it(1-\hat{P})L}$ on a non-zero $\rho_{ir}(0)$ could also be such that the term disappears.

2. **The Markovian approximation** requires that $\hat{G}(t')$ decreases to zero over a certain timescale, the 'relaxation time', τ. Thus, for times t' greater than the relaxation time τ, $\hat{G}(t') = 0$. Furthermore, it requires that ρ_r does not vary much over this timescale τ, and so $\hat{G}(t')$ drops off more rapidly than the timescales over which ρ_r evolves. To sum up: the *key physical idea* of the Markovian approximation is that there is a relaxation time τ over which the integral kernel drops off and over which ρ_r does not vary much (Wallace, 2015b, p. 292).[7]

   Provided that these physical features hold, then the following mathematical moves can be made:

   a) If the integral upper limit t is greater than τ extending the integration interval to ∞ makes no difference to the value of the integral; $\int_{t_0}^{\infty} dt'\hat{G}(t')\rho_r(t-t') \simeq \int_{t_0}^{t} dt'\hat{G}(t')\rho_r(t-t')$.

   b) If ρ_r varies very slowly over τ, $\rho_r(t-t') \approx \rho_r(t)$ for $t' < \tau$. (If $t' > \tau$ this approximation does not hold, but since $\rho_r(t-t')$ is multiplied by $\hat{G}(t')$ which is ≈ 0 for $t' > \tau$, we can replace $\rho_r(t-t')$ by $\rho_r(t)$.)

   c) Thus, if the Markovian approximation holds, we can replace the second term $\int_{t_0}^{t} dt'\hat{G}(t')\rho_r(t-t')$ of equation (3.5) by $\int_{t_0}^{\infty} dt'\hat{G}(t')\rho_r(t)$.

Provided that the initial state assumption and the Markovian approximation hold, we thus arrive at an autonomous equation —*the master equation*— for the relevant degrees of freedom, ρ_r:

[7] This general assumption of 'different timescales' is of course used much more widely than just in the ZZW framework. For example, in Reif (2009, Ch. 14) the derivation of the Boltzmann equation requires a similar assumption: that $f(\vec{r}, \vec{v}, t)$ does not vary appreciably during time intervals of the order of the collision time, nor over spatial intervals of the order of intermolecular forces.

I now offer an intuitive explanation of the general situation, by extending Zeh's discussion using a metaphor of a forest. Within the irrelevant information, Zeh (2007, p. 65) distinguishes the 'doorway' from 'deep states', which are in different 'channels'. So there are three 'channels': (A) 'relevant', (B) 'doorway', (C) 'deep' — and these are analogous to (A) a clearing in a wood, (B) the sunny woodland surrounding the clearing, (C) the dark woods. Zeh gives the following example: (A) is a one-particle marginal density, (B) encodes two-particle correlations and (C) encodes three-or-more particle correlations (cf. the BBGKY hierarchy). Now, the non-Markovian term in the pre-master equations gives the contributions to $\frac{\partial \rho_r(t)}{\partial t}$ at t from the part of ρ_r that became irrelevant at $t - t'$ and remained irrelevant until time t: at which point it becomes relevant again.

This 'information becoming irrelevant' is like people in the clearing (A) wandering into the sunny woodland (B). Thus, the relaxation time τ is the time taken for the people who arrived in the sunny woodland (B) to wander either back to the clearing (A) or into the dark woods (C). The key assumption is that once in the dark woods (C), no one can find their way back to the clearing (A) again. In less picturesque terms: the three-or-more particle correlations are not dynamically relevant for the one-particle marginal density.

This metaphor also encompasses the famous recurrence theorem. If you wander around a (finite) woodland for an incredibly long (i.e. recurrence) time, you will eventually find your way back to the clearing. As I will discuss in Section 3.3.1, on recurrence timescales the 'deeply' irrelevant states (C), e.g. three-or-more particle correlations, become relevant (A) again.

$$\frac{\partial \rho_r(t)}{\partial t} \approx \hat{D}\rho_r(t), \tag{3.6}$$

where $\hat{D} := \int_{t_0}^{\infty} dt' \hat{G}(t')$.

This completes Stage 3. For our purposes, there are three comments to make.

(1) This schematic equation (3.6) has specific forms for specific systems (Penrose, 1979, p. 1986); "various particular cases of it include the (empirically verified) equations of decoherence, of radioactive decay, and of diffusion and equilibration in dilute gases" (Wallace, 2015b, p. 292).

(2) We can now describe the irreversible behaviour using a generalised version of the Gibbs coarse-grained entropy. The coarse-grained Gibbs entropy S_{cg} (in equation 3.1) can be written as a functional of ρ and $\hat{P}_{cg}$:

$$S_{cg}[\hat{P};\rho] = -k_B \int \hat{P}_{cg}\rho(q,p) \ln \hat{P}_{cg}\rho(q,p) d^{3N}q d^{3N}p. \tag{3.7}$$

And similarly more generally: we define, for any ZZW projection $\hat{P}$, obeying equations (3.5) and (3.6), the entropy:

$$S[\rho_r] := S[\hat{P};\rho] := -k_B \int \hat{P}\rho(q,p) \ln \hat{P}\rho(q,p) d^{3N}q d^{3N}p. \tag{3.8}$$

This quantity *can* increase — like S_{cg}, as noted after equation (1). Thus Zeh writes: "if $\hat{P}$ only destroys information, the master equation describes never-decreasing entropy" (Zeh, 2007, p. 65):

$$\frac{dS[\rho_r]}{dt} \geq 0. \tag{3.9}$$

For a proof, see Tolman (1938, p. 171), Huang (1987, p. 74), Reif (2009, p. 624) and for the quantum context, see Landsberg (1990, p. 145).

(3) Finally, and most importantly for our interests: this closed equation 3.6 is *irreversible* (Zwanzig, 1960, p. 1340).

3.3 Why does this method work?

Why does the ZZW framework lead to empirically successful equations? This success is surprising because, after all, the coarse-graining projection (and the ensuing C^+ dynamics) cannot be implemented by the "official" microdynamics. Given Liouville's theorem, the microdynamics of the closed system cannot really cause the velocity correlations to be erased (in the case of the Boltzmann equation), or really delete the off-diagonal density matrix elements (in the case of the Pauli master equation). In short: the TRI microdynamics of the closed system cannot dynamically implement the coarse-graining projection.

In order to explain the success of irreversible equations in SM there have been three broad strategies:

- (1) Interventionists, e.g. Bergmann and Lebowitz (1955); Blatt (1959); Ridderbos and Redhead (1998), argue that perturbations from the environment cannot be neglected. Thus, the system cannot be treated as closed. (In the ZZW terminology, the environment dynamically implements the projection, so that ρ_r, rather than ρ, is the correct description of the subsystem.)
- (2) Others advocate changing the underlying microdynamics so that the coarse-graining projection is dynamically implemented. Albert (2000) and Prigogine and Stengers (1984) advocate non-TRI microdynamics in the quantum and classical case respectively. (In the ZZW terminology, the non-TRI dynamics yields $\rho \mapsto \rho_r$.)
- (3) Some, such as Wallace (2012a), propose that under special conditions the irreversible SM dynamics will give the same density over the relevant degrees of freedom as the microdynamics.

For the remainder of the chapter, I only focus on the third of these strategies, which I call 'the special conditions' account. In Section 3.3.1, I consider this account and the required 'meshing' condition. Section 3.3.2 considers when a density satisfies this condition and reports Wallace's proposal. This will lead into the idea of a 'Past Hypothesis'; (although, as mentioned in Stage 2 of Section 3.2, an in-depth discussion of the controversial Past Hypothesis is beyond the scope of this chapter).

3.3.1 The special conditions account

The third strategy claims that under certain conditions the microdynamics will induce the same probabilities for the relevant degrees of freedom, as the C^+ coarse-grained dynamics governing ρ_r. On this view, the generalised coarse-graining projection is not *dynamically* implemented. Thus, ρ and ρ_r are two distinct densities.

How do we find ρ_r at a given time T? There are two "routes". As discussed in Section 3.2, the C^+ dynamics for a period $t_0 < t < T$ is defined by evolving the density by the microdynamics $\hat{U}$ for a very short time Δt, then applying the projection $\hat{P}$, then evolving under $\hat{U}$ for Δt, then $\hat{P}$... etc. This means that irrelevant details are thrown away at every step. In contrast, the Liouvillean microdynamics $\hat{U}$ evolves the full density ρ for the period $t_0 < t < T$; and then one finds the relevant part of the density by applying $\hat{P}$ at T; so on this "route", coarse-graining occurs only once at the end of the time-period. Thus, the condition that these two different dynamics give the same density $\rho_r(T)$ can be expressed by the diagram in Figure 3.2 commuting.

$$\begin{tikzcd}
\rho_r(t_0) \arrow[r, "C^+(t)"] & \rho_r(t) \\
\rho(t_0) \arrow[u, "\hat{P}"'] \arrow[r, "U(t)"] & \rho(t) \arrow[u, "\hat{P}"']
\end{tikzcd}$$

Figure 3.2: ρ and $\hat{P}$ are forwards-compatible if the two routes to $\rho_r(t)$ give the same answer.

Following the terminology suggested by Wallace (2011), let us call those states ρ for which diagram 3.2 commutes **forwards compatible** with coarse-graining $\hat{P}$. So forwards compatibility means that it does not matter whether you coarse-grain at every time step Δt or just once, at the end. Note that forwards compatibility is relative to a particular choice of coarse graining $\hat{P}$. Thus this is a condition of 'harmony' between the evolution of ρ and the coarse graining $\hat{P}$. For example, had the size of the coarse-graining boxes ΔV_m averaged over in Gibbs' original example been chosen to be very large, then ρ might well not be forwards-compatible with this coarse-graining, $\hat{P}_{cg}$. In the wider literature on inter-theoretic relations, such a forwards-compatible scenario is sometimes described as 'meshing' dynamics (e.g. Butterfield (2012), List (2016)).

However, we cannot expect harmony to "reign supreme". Not all densities ρ will satisfy Figure 3.2's meshing condition: Loschmidt's reversibility objection vividly reminds us that if we were to reverse the momenta of the components of a fibrillating ink drop, it would coalesce back in a manner incompatible with the 'smooth-spreading out' coarse-grained dynamics. (More specifically: the time-reverse of a density ρ initially forwards-compatible and on a trajectory of increasing entropy will not itself be forwards-compatible.)

And due to Poincaré's recurrence theorem, nor will any ρ satisfy the meshing condition for all time. (More specifically: in the ZZW framework, recurrence implies that the integral kernel $\hat{G}$ in equation (3.5) must increase again so that at the recurrence time it has returned to its original value. Therefore the upper limit of the integral in the Markovian approximation strictly cannot be taken to ∞, but at most to some large — but sub-recurrent — time T. Consequently, the Markovian approximation is only valid for sub-recurrent times.)

3.3.2 When is a density forwards-compatible?

Characterising those densities ρ which are forwards-compatible is a harder job than ruling out candidate densities. A density ρ will be forwards-compatible provided that the density ρ_{ir} (and the details such as correlations encoded in it) that are thrown away by $\hat{P}$ do not matter for the forwards-evolution of ρ_r. One clear case where this is *not* true is Hahn's spin-echo experiment (Hahn, 1950). The application of a radio-frequency pulse causes dephased spins (precessing in a magnetic field) to realign and thus emit an 'echo' signal (for a recent philosophical discussion see Frigg (2010, §3.5.1)). The correlations — that are ignored from the coarse-graining perspective — are crucial for the later 'echo signal'. Indeed, the spin-echo experiment has been described as a 'Loschmidt demon' which reverses the velocities $v \mapsto -v$.[8]

Given the above discussion of the Loschmidt reversibility objection, here too the density ρ is clearly not forwards-compatible. Consequently, the spin-echo is not a *surprising* counterexample to the coarse-graining framework — which we can only expect to be successful when Figure 3.2 commutes: i.e. when the information (in this case, correlations) thrown away by the coarse-graining projection $\hat{P}$ are not crucial — unlike the spin-echo case.

Ridderbos and Redhead (1998, p. 1237) and Blatt (1959, p. 749) generalise from the spin-echo case to reject the coarse-graining framework altogether.[9] However, rather than claiming that the spin-echo case reveals coarse-graining to be empirically inadequate, it seems fairer to say the density ρ is patently not forwards-compatible and so we do not expect coarse-graining methods to apply.[10]

Naturally, the following question arises: why should we expect the spin-echo ('correlations-are-crucial') type of case to be the exception rather than the rule? To this, the reply can only be that 'nature is kind': often — i.e. in the irreversible equations of SM — ρ_{ir} *is* irrelevant for the evolution of ρ_r.

Nonetheless, one might ask: what informative condition can be used to pick out the forwards-compatible scenarios? Since the presence of 'crucial correlations' was the problem in the spin-echo case, perhaps removing them is the answer: ensuring there is no irrelevant information at all is one way to avoid the failure of compatibility. Indeed, this is what the initial state assumption $\rho_{ir}(t_0) = 0$ in Section 3.2 achieved — and alongside the Markovian approximation, this was used to construct the C^+ dynamics. In similar vein, Wallace stipulates that *'Simple'* initial densities ρ will not have crucial conspiratorial correlations encoded in their irrelevant degrees of freedom; he defines "a Simple distribution as any distribution specifiable in a closed form in a simple way without specifying it as the time evolution of some other distribution" (Wallace, 2011,

[8] More accurately, the spin's velocities are unaltered, but the order of the spins is altered by reflection in the x-z plane. However, "the grain of truth in the standard story is that a reversal of the ordering with unaltered velocities is in a sense 'isomorphic' to a velocity reversal with unaltered ordering" (Frigg, 2010, p. 64).

[9] Blatt concludes that "it is not permissible to base fundamental arguments in statistical mechanics on coarse-graining" (Blatt, 1959, p. 749).

[10] Lavis (2004, p. 686) further defends coarse-graining.

p. 19).[11]

Note, however, that such a condition — the initial state assumption or Wallace's Simplicity condition — can only be applied once.[12] The initial state A in Figure 4.1 — confined to four Gibbsian cells, or, in the analogy, the ink drop's initial state — is Simple (or equivalently it satisfies the ZZW initial state assumption). However, it then fibrillates over the available phase space and thus is no longer Simple. Whilst initially at t_0 there was no irrelevant information, this is no longer the case: $\rho_{ir}(t_1) \neq 0$. Yet — we hope! — nonetheless $\rho(t_1)$ is still forwards-compatible. Accordingly, the 'Simple' states are a subset of the forwards-compatible states. Thus, given the microdynamics, imposing such an initial condition is a sufficient but not necessary initial condition for ensuring that ρ is forwards-compatible. The plausibility of such initial conditions on probabilities densities will depend on one's interpretation of probability in SM. 'Simplicity' fits especially well with a Jaynesian account: Jaynesians interpret ρ as encoding our ignorance of the system. If all we know is the system's macrostate, then claiming that ρ is uniform across this state ensures that ρ is Simple.

Given that such an initial condition can only be applied once, when should we apply it? Practising physicists apply it at the beginning of the time of interest, t_0 (option 1). But this leads to a problem akin to that facing Boltzmann's combinatoric argument. By parity of reasoning, this licences the construction of the C^- dynamics prior to t_0, and the C^- dynamics yields anti-thermodynamic trajectories prior to t_0. Such parity problems motivate the 'Past Hypothesis'; in the Boltzmannian case that the initial macrostate of the universe had a 'low entropy' (Albert, 2000, Ch. 4). Here, this parity problem motivates Wallace (2011, p. 22) to apply the initial state assumption to the beginning of the universe (option 2). An in-depth analysis of the Past Hypothesis —and the different possible forms it could take cf. Wallace (2011)— is not possible here, but I can allay one worry: provided Markovian approximation holds good, the choice between applying this condition in the manner of physicists (option 1) and a Past Hypothesis (option 2) will not lead to dramatic empirical differences. The difference between t_0 for options 1 and 2 is dramatic: 13.7 billion years. One might think that this should lead to equally dramatic differences in the constructed equations, as t_0 appears in the premaster equation. And thus one might hope to adjudicate between options 1 and 2 on these *empirical* grounds. But the key physical insight behind the Markovian approximation explains why despite the dramatic difference between t_0 for options 1 and 2, there

[11] One might object that this definition is vague. Instead consider this 'Simplicity' condition as: an overarching condition to capture what is similar across those densities which satisfy the initial state assumption for different $\hat{P}$s. A given ρ satisfying the initial state assumption at t_0 will ensure that – with respect to a given $\hat{P}$ and thus a given definition of 'irrelevant' — ρ is 'Simple' at t_0. But of course there are many densities that count as 'Simple' in some sense, but not in the respect required for the initial state assumption ($\rho_{ir}(t_0) = 0$) for a particular $\hat{P}$.

[12] Wallace points out that it would be excessive to apply it more than once: the microdynamics are deterministic and so fixing ρ at one time fixes ρ for all times.

need be an accompanying dramatic empirical difference, as follows. Provided that the Markovian approximation holds good and —as is uncontroversial— the recurrence time is much much greater than 13.7 million years, if we apply $\rho_{ir} = 0$ at the beginning of the universe, there will not be 13.7 billion years' worth of 'irrelevant information' (e.g. correlations encoded in ρ_{ir}) that is liable to be about to become dynamically relevant for ρ_r (and so causing empirical differences between option 1 or 2). The only potential difference will be the information τ seconds ago. (See Zeh (2007, p. 64) for more details).

In summary: When the forwards-compatibility condition fulfilled, C^+ dynamics gives the same values for relevant ρ_r as the microdynamics. Not all densities ρ are forwards-compatible and nor is any density forwards-compatible for all times: as shown by the reversibility and recurrence objections respectively. When considering how to determine whether a given ρ is forwards-compatible or not, one suggestion was that a probability density will be forwards-compatible if it satisfies the initial state assumption at t_0 (or in Wallace's terminology is 'Simple' at t_0). However, whether t_0 should be taken to be at the beginning of time of interest (option 1) or the beginning of the universe (option 2) is a contentious matter.

3.4 Anthropocentrism and illusion: two objections

If coarse-graining is empirically successful (as I have claimed) then perhaps no further justification is required. This would be a tempting line to take, were it not for the literature's containing a barrage of criticisms of coarse-graining. For example: coarse-graining 'seems repugnant to many authors' (Uffink, 2010, p. 197) and is even claimed to be 'deceitful' (Redhead, 1996, p. 31). The coarse-grained time-asymmetry is also called 'illusory' (Prigogine, 1980) and potentially 'subjective' (Denbigh and Denbigh, 1985, p. 53).

This purported subjectivity of coarse-graining leads to concerns about the status of the time-asymmetry. According to Davies, 'it is indeed a matter of philosophy rather than physics to decide if the coarse-grained asymmetry is 'real' or not' (Davies, 1977, p. 77). Furthermore, the potentially unusual or subjective status of the coarse-grained asymmetry within physics leads Grünbaum (1973) to discuss whether scientific realism is incompatible with coarse-graining approaches in SM. More broadly, determining this status of the asymmetry is part of a wider philosophical project of untangling 'what is genuinely an aspect of reality from what is a kind of appearance, or artifact, of the particular perspective from which we regard reality' (Price, 1996, p. 4).

Summing up, it seems to me that these objections can be divided into two camps:

(Illusory): First, the asymmetry is a mere artifact of coarse-graining and so is *illusory*.

(Anthropocentric): Secondly, it arises from our perspective and so is *anthropocentric*.

Given these concerns and objections, coarse-graining requires some conceptual, not just empirical, justification. I propose that this task can be split into two:

(Choice): What is the justification for the *choice* of coarse-graining projection?

(At all): Why is it legitimate to coarse-grain *at all*?

A justification for coarse-graining may of course purport to answer both questions. And the answers may be linked. For example, if the justification for the choice of coarse-graining projection was deemed to be unacceptably subjective, then this might lead one to believe that coarse-graining *at all* is unacceptable. However, the two issues can also come apart. For example, a justification for coarse-graining might only motivate why it is an acceptable procedure in general, but remain silent on how to choose a particular coarse-graining projection.

In Section 3.4.2, I will consider and reject the 'measurement imprecision' justification and discuss how it lies behind the *(Illusory)* and *(Anthropocentric)* objections: before giving, in Section 5, my favoured justification. But first, I consider the two objections in more detail — in Section 3.4.1.

3.4.1 The two objections in more detail

The claim that the coarse-grained asymmetry is an "illusion" (Prigogine, 1980): as cited in Denbigh and Denbigh (1985, p. 56) is rooted in the action of $\hat{P}$. The contention is that $\hat{P}$ 'distorts' ρ and the gap between ρ and ρ_r is the source of the coarse-grained asymmetry. Every time we apply $\hat{P}$ we edge away from the correct density ρ — in particular we edge away from the correct value of the Gibbs (fine-grained) entropy by a certain amount: "the required increase in the coarse-grained entropy is obtained by disregarding the dynamical constraints on the system" (Ridderbos, p. 66). By repeatedly coarse-graining (as is done in the C^+ dynamics), we generate the coarse-grained asymmetry. "The repeated coarse-graining operators appear to be added 'by hand', in deviation from the true dynamical evolution provided by U_t" (Uffink, 2010, p. 197). That is, the coarse-grained asymmetry exists merely in virtue of the continual coarse-graining in the C^+ dynamics — each coarse-graining increases S_{cg} by some small amount so that eventually an asymmetry is produced. "Perhaps most worrying, the irreversible behaviour of S_{cg} arises almost solely due to the coarse-graining" (Callender, 1999, p. 360). Thus, since the asymmetry stems from the infidelity of coarse-graining, it is illusory.

This $(Illusory)$ objection has the following form:

- P1. The action of $\hat{P}$ is to deliberately distort the correct density ρ.

- P2. The asymmetry only arises from the repeated coarse-graining every Δt in the C^+ dynamics.
- Conclusion: The coarse-grained asymmetry is an illusion.

Next I consider the $(Anthropocentric)$ objection. According to this objection, the coarse-grained asymmetry, in particular the coarse-grained entropy, is not an objective physical quantity, like energy or mass but rather is 'agent-centric'. For example, Wigner and Jaynes have called entropy 'anthropocentric' (Jaynes, 1965). The terms 'subjectivity' and 'anthropocentrism' are used interchangeably in this debate. Denbigh and Denbigh (1985) helpfully distinguish two kinds of objectivity (and thereby of subjectivity). *Objectivity*$_1$ is intersubjective agreement. *Objectivity*$_2$ is stronger. It requires the phenomena in question to be independent of human cognition. In the debate about coarse-graining, intersubjective disagreement is not the issue. Rather it is the second kind of subjectivity ($\neg$*Objectivity*$_2$) that is at stake, which I earlier dubbed 'anthropocentrism'.

The reason for this charge of anthropocentrism is as follows. In the case of the archetypal Gibbsian coarse-graining $\hat{P}_{cg}$ the size of the boxes is chosen by us. "There are no laws of physics which determine the size of the [cells]" (Denbigh and Denbigh, 1985, p. 51): merely our preference determines the choice. Furthermore, "the increase of entropy and the approach to equilibrium would thus apparently be a consequence of the fact that *we* shake up the probability density repeatedly in order to wash away all information about the past, while refusing a dynamical explanation for this procedure" (Uffink, 2010, p. 196). In addition, the partition is *chosen* by us: "the occurrence and direction of a temporal change of the entropy... depends essentially on *our human choice* of the *size* of the finite equal cells of boxes into which we partition... phase space " (Grünbaum, 1973, p. 647, emphasis in original). The objection extends to all instances of $\hat{P}$; "a Zwanzig projection (describing generalized coarse-graining) can be arbitrarily chosen for convenience" (Zeh, 2007, p. 67).

Grünbaum (1973) points out that the charge of anthropocentrism here differs from the more general claim that scientific theories are human constructs. It seems that the Standard Model could describe the world, even if there were no (human) observers. Yet, according to the *Anthropocentric* critique, this would not be the case for entropy, and the coarse-grained description.

Lying behind these objections is a particular justification of coarse-graining: the measurement imprecision (MI) justification: to which I now turn.

3.4.2 Against the justification by measurement imprecision (MI)

In the literature, the most common justification for coarse-graining is that our measurements have limited precision. “The coarse-graining approach makes essential use of the observation that we only have access to measurements of finite resolution” (Ridderbos, 2002, p. 66). Thus, we can never locate a system precisely in phase space; we only know p and q to a certain degree of accuracy. The cells over which we average with the $\hat{P}_{cg}$ for the archetypal Gibbsian coarse-graining have a size which corresponds to “the limits of accuracy actually available to us” (Tolman, 1938, p. 167). Because we could never, *ex hypothesi*, measure the system accurately enough, we are unable to distinguish between the coarse and fine-grained distributions ρ and ρ_r. Thus, according to this measurement imprecision (MI) justification, the answer to (*Choice*) is that we must pick the coarse-graining $\hat{P}$ that matches our observational capacities. For those coarse-grainings $\hat{P}$ whose selection is justified by the indistinguishability between ρ and ρ_r, the MI justification also answers why (for those particular projections) coarse-graining (*At all*) is justified — because we cannot tell the difference.

Appealing to appearances originates from Gibbs’ ink analogy. Whilst the ink drop’s volume is constant, it fibrillates throughout the water, and so it *appears to us* to be uniformly distributed. Our limited powers of observation cannot distinguish between the fibrillated case and the locally uniform distribution resulting from coarse-graining.

A similar argument arises in the Boltzmannian approach to SM, where phase space is partitioned into ‘macrostates’. Every microstate corresponds to one macrostate. A particular macrostate is defined by values of macrovariables, such as volume, temperature and pressure. These macrostates are sets of microstates that are ‘empirically indistinguishable’. Thus, an appeal is once again made to our observational capacities.[13]

The $(Illusory)$ and $(Anthropocentric)$ objections arise from this justification of coarse-graining (rather than coarse-graining itself). The claim that the coarse-grained asymmetry is illusory is bolstered by the MI justification, since it implies that if we were to be able to measure the system more precisely (in the idiom of Gibbs’ analogy to *see* the thin fibrillating tubes of ink rather than the smooth spreading) then the asymmetry would disappear. The coarse-grained asymmetry would thus be an illusion stemming from the imprecision of our measuring devices. The claim that the asymmetry is anthropocentric is also underwritten by the MI justification. If the coarse-grained ρ distribution is indistinguishable from the fine-grained ρ distribution *to us* and thus the choice of $\hat{P}$ depends our capabilities, then the asymmetry would be anthropocentric.

However, the MI justification is unsatisfactory. This is not (only) because it leads to the illusion and anthropocentric objections, but also, even on its own terms: it is both insufficient and unnecessary for justifying coarse-graining. (However, other purposes

[13] But the Boltzmannian partition is not necessarily the same as the Gibbsian cells.

for which measurement imprecision may be important will be briefly discussed in Sections 3.7 and 3.8.1).

The imprecision of our measurements is not a *sufficient* justification for implementing a coarse-graining projection $\hat{P}$, since choosing a projection that fits with the limits of observation will not always lead to autonomous irreversible dynamics of the type given by the ZZW framework. "Observability of the macroscopic variables is not sufficient... It is conceivable (and occurs in practice) that a particular partition in terms of observable quantities does not lead to a Markov process" (Uffink, 2010, p. 196). That is, a coarse-graining could reflect our measurement precision but not lead to an example of useful dynamics: in particular, to autonomous C^+ dynamics. Therefore, measurement imprecision is not sufficient for answering *(Choice)*.

Furthermore, appealing to measurement imprecision is not *necessary* for explaining why we should choose any particular coarse-graining $\hat{P}$. If it were, we would in every case have to ascertain the imprecision of particular measuring devices and accordingly choose a coarse-graining $\hat{P}$. Yet, in Section 3.2, this is not how coarse-graining projections were chosen; and the details of particular measuring devices (or the resolution of our eyes) are in fact never used in constructing equations in the ZZW framework. It seems unlikely that advances in the science of microscopy will lead to different choices of $\hat{P}$.[14]

Thus appealing to the limited precision of our measurement devices is incapable of justifying the choice of coarse-graining projections (*Choice*). The MI justification only answers (*At all*) in virtue of answering *(Choice)* in particular cases, and thus its failure to answer to *(Choice)* means that it automatically does not answer (*At all*). With MI thus rebutted, I now outline my proposed alternative justification.

3.5 An alternative justification

Applying $\hat{P}$ throws away details. Why would throwing away details ever be a good move? One motivation for moving to the coarse-grained description is that modelling the evolution of ρ under the Liouvillean dynamics is computationally intractable, because solving the equations of motion for some 10^{23} particles is infeasible.

Were this the only motivation for coarse-graining, one might be misled into believing that in an ideal world where we were equipped with a sufficiently powerful computer and the initial states of each of 10^{23} particles, the coarse-grained description would be

[14] My rebuttal of the necessity of the MI justification takes its proponents at their word. But perhaps this is uncharitable, for in reality, typical discussions of the construction of autonomous equations are often schematic— they merely assume there is such a projector that satisfies the required assumptions, without a detailed demonstrations that the projector does indeed fulfil these assumptions. As I have not investigated such demonstrations, it is an open question whether for that project —rather than the construction of autonomous dynamics— measurement imprecision is necessary.

dispensed with. Yet something would be lost, if upon receiving all the information and extraordinarily powerful computers, we ditched the discipline of SM: and this reveals a general point about the assumptions in SM. Namely: as I argue in Section 3.5.1, computational intractability is not the only motivation for such approximations and assumptions. In Section 3.5.1, I distinguish between Galilean idealisation and abstraction, and then classify coarse-graining as abstraction to a higher level of description. This, plus the desideratum that the dynamics at this level be autonomous, allow me to justify coarse-graining. Then, in Section 3.5.2, I illustrate these ideas of abstraction and autonomy with the Game of Life.

3.5.1 Abstraction and autonomy

There are two reasons that such leaps in our computational capacity would not make SM 'superfluous'. Firstly, it is unclear in what sense solving some 10^{23} coupled equations would constitute an *explanation* of the behaviour of the gas.[15] Secondly, a statistical mechanical system such as a gas "exhibit[s] perfectly definite regularities in its behaviour" (Tolman, 1938, p. 2). Such regularities would be lost amongst the morass of detail at the fundamental (or lower) level. This difference in levels of description is particularly vivid in the case of coarse-graining; by moving to the lower-level Liouvillean dynamics, we not only lose explanatory power but also some very useful equations that determine transport coefficients and relaxation times.

At this point, we need to distinguish different strategies for simplifying scientific descriptions. This is a large topic and the words at issue —idealisation, abstraction and approximation— are terms of art that different authors construe differently, but I will crudely categorise strategies as Galilean idealisations or abstractions. A Galilean idealisation introduces 'deliberate distortions' (Frigg and Hartmann, 2006), familiar from the standard examples of frictionless planes and perfectly rational economic agents. A common way to think about such idealisations is by analogy to a perturbative series. The behaviour of the target system is veridically described by the full series, but a successful idealised description is akin to the first term of the series. Adding the higher-order terms renders the idealised description more accurate and furthermore, explains the success of the idealisation even if these terms are not actually calculated (Batterman, 2009, p. 17). Often Galilean idealisations are used in order to render a problem more tractable — and in an ideal world, we would remove the idealisation (and so add all the terms of the series in) — and this would lead to a more accurate representation.[16]

[15] Some might argue that whilst the solution of the 10^{23} equations might not be the best explanation, it is nonetheless *an* explanation. However, the details of the vast debate about explanation are not needed for what follows.

[16] In Weisberg's terminology, the 'representation ideal' would be to *remove* the idealisation (Weisberg,

In contrast, I take abstraction to be the omission, or throwing away, of certain pieces of information (Knox, 2016; Thomson-Jones, 2005). This corresponds to a broad category in the literature: Weisberg (2007)'s minimal modelling, Cartwright's abstraction and Aristotelian idealisation (Frigg and Hartmann, 2006). This category involves 'throwing away details, stripping away, keeping only the core causal factors'.[17]

Thus I claim: coarse-graining is not a Galilean idealisation. If it were, there would be certain details those inclusion would *improve* the coarse-grained description. Yet, in the ZZW framework, this is not so. Indeed, we know *exactly* which details would need to be added to render a more complete description — the information about the irrelevant degrees of freedom that we threw away! But clearly if we were to add ρ_{ir} back in, we would no longer have the coarse-grained, and useful, equations found in Section 3.2.

Instead, coarse-graining is abstraction. ρ_r omits irrelevant information, which has been discarded by $\hat{P}$. For instance, in the archetypal Gibbsian case, the action of the coarse-graining projection $\hat{P}_{cg}$ is to omit exactly how the probability varies across the coarse-graining cell as only the probability of the entire cell is relevant: "how full the cell is, rather than how it is filled". Some projections take a density in a given equivalence class to be an exemplar of that class (Wallace, 2011, p. 9). In such cases, only the fact that the density is in the equivalence class is relevant, not which member of the class it is.[18] In the case of $\hat{P}_{\mu}$, information about the correlations between particles is omitted.[19]

Thus ρ_r is a new variable germane to this higher-level of description implicitly defined by a given $\hat{P}$: rather than a distorted replacement of ρ, which is how an idealisation conception of coarse-graining would interpret ρ_r. As ρ_r forms part of a higher-level of description it need not be in tension with ρ, just as descriptions in biology need not be in tension with descriptions in psychology. (In the terminology of chapter 1, these descriptions are about different subject matters). ρ_r is not an 'idealised' version of ρ containing false elements, as omission need not get in the way of telling a true causal story (Strevens, 2008; Lewis, 1986a). Thus, coarse-graining *at all* is justified because it allows us to abstract to a higher-level of description. This is my proposed answer to Section 3.4's *(At all)*.

2007, p. 642).

[17] Of course, there are other categorisations - e.g. McMullin (1985) has six types of idealisation. Norton (2012) discusses approximation and idealisation in a different sense. Notably, different $\hat{P}$s might be (sub)-categorised differently according to a more fine-grained classification. However, all that matters here is that coarse-graining is not a Galilean idealisation.

[18] To link to chapter 1: clearly this is an example of one of List's abstraction maps, σ.

[19] Of course, adding in *some* correlations (i.e. adding in the third tier of the BBGKY hierarchy) may lead to a more empirically successful autonomous equation than the autonomous equation describing the evolution of the one-particle marginal density. Indeed, e.g. doing so can provide corrections to the Boltzmann equation. But of course, we wouldn't want to add in *all* the correlations in the BBGKY hierarchy — in the limit, doing so would take us back to the reversible dynamics. And indeed, truncation at first or second equations in the hierarchy is the key benefit of the BBGKY approach — it is useful *because* we can get away with only considering the lower hierarchy.

$\hat{P}$ abstracts to a higher level of description. Yet we don't *just* want to abstract to a higher level: we want a *theory* of the goings-on at this level. For example, suppose $\hat{P}_{cam}$ coarse-grains the position and mass distribution of people in Cambridge to the centre of mass of this population. The information about the masses and locations of individuals has been thrown away, leaving a more abstract description of the population. However, discussing the centre of mass of Cambridge's population is not going to be useful, if the only way to find out how this centre of mass moves is to consider the movement of all the individuals and then re-average. If we cannot say anything about what is going on a higher level of description without invoking information from the lower level, then the higher level of description is not going to be useful.[20]

But not having to refer to the lower-level details in describing the goings-on at the higher level of description is precisely what the *autonomy* condition in the ZZW framework captures. Recall that the dynamics are autonomous if they were of the form $f(\rho_r)$ rather $f(\rho_r, \rho_{ir})$; the dynamics for the relevant degrees of freedom have no functional dependence on ρ_{ir}. In other words, ρ_{ir} is not a 'difference-maker' for the evolution of ρ_r (Woodward, 2005; Strevens, 2008, Ch. 3). Note, however, that whilst the idea of different descriptions is contained in the concept of autonomy, no notion of hierarchy is implied. There could be different descriptions without one being 'higher' than another (cf. List and Pivato (2015, p. 150, fn. 41). Thus the 'higher-level' aspect of this justification comes from taking $\hat{P}$ as abstracting from irrelevant details. The terminology of 'relevant' and 'irrelevant' degrees of freedom is highly appropriate; for if the dynamics weren't autonomous then the so-called 'irrelevant' details would indeed be relevant.

Now, by taking this cue from the ZZW framework, it is clear what justifies the choice of any particular coarse-graining map. Whilst any coarse-graining map can be used to find a pre-master equation, not every $\hat{P}$ will lead to coarse-grained irreversible dynamics. Only those coarse-grainings of a system that satisfy the two conditions (in Stage 3 of Section 3.2) will lead to autonomous dynamics.[21] Thus, the choice of coarse-graining map is determined by whether it results in successful C^+ dynamics. I agree that this criterion will not help physicists discover new, useful maps. The class of successful $\hat{P}$s will not look especially unified. But this is to be expected; each case requires details of the particular system at hand. Thus as Uffink (2010, p.195) says: "it is 'the art of the physicist' to find the right choice, an art in which he or she succeeds in practice by a mixture of general principles and ingenuity, but where no general

[20] List and Pivato (2015, p. 135) go further. In their framework, the lower-level language is by definition unavailable at the higher level.

[21] Autonomy in the sense of 'not referring to ρ_{ir}' is achieved by the initial state assumption. But for autonomy in the sense of not depending at all on t, the Markovian approximation needs to be satisfied.

guidelines can be provided".[22]

To summarise, this alternative justification answers Section 3.4's two justificatory questions as follows:

- ⋆ *(Choice)* - The choice of a particular map is determined by the desideratum of finding autonomous dynamics.
- ⋆ *(At all)* - Applying a map $\hat{P}$ abstracts to a higher level of description.

3.5.2 An illustration: the Game of Life

The key ideas of autonomy and abstraction are vividly illustrated by Conway's Game of Life: a standard example of the complexity science, and emergence, literature (see e.g. Bedau and Humphreys (2008, Ch. 8,9,11,16,17)). The Game of Life is a cellular automaton that operates via a simple rule: at each time-step, whether a cell of the Grid is ON or OFF is determined by how many of its eight neighbours are ON. Despite the extreme simplicity of the dynamical rule, a rich variety of patterns can evolve in the grid. These stable shapes have characteristic movements and so are given vivid names: glider guns spawn gliders moving across the grid, eaters destroy other shapes they 'encroach' on, and puffer trains move across the grid leaving behind debris in their wake — to name but a few. Whilst the sheer variety of the Game of Life cannot be easily conveyed in words (and is best appreciated by viewing a video of the evolution of a Life Grid), to give an idea of the complexity that can arise: the Universal Turing machine has been constructed in the Life Grid (Poundstone, 2013, p. 213).

When discussing the Life Grid, we can abstract to a higher level of description and, as done above, describe the goings-on in terms of the menagerie of 'gliders' and 'blinkers' rather than in terms of the cells. For example, the glider moves across the grid with velocity $c/4$, where c is the 'speed of light' (in the sense of being the 'speed limit' — this maximum speed is one cell per unit time). This alternative description of gliders "has its own language, a transparent foreshortening of the tedious descriptions one could give at the physical level" (Dennett, 1991, p. 39). Discussing the gliders' motion in this way is predictively successful. Furthermore, often these descriptions are autonomous: we need not keep referring back to the lower-level, i.e. cell-level, details.[23] But, of course, theoretically we could have calculated the evolution of the grid at the cell-level and then, at the end, abstracted to the higher-level, e.g. glider-level, of description.

[22] Of course, in individual cases, there will be the further explanatory project of showing that the two required assumptions are satisfied by a chosen $\hat{P}$ — and this will give us further insight into why in these particular cases autonomous dynamics are possible, i.e. why our desideratum is fulfilled. But as a *general* answer to (*Choice*) — the only rationale for picking any $\hat{P}$ is that it leads to an autonomous dynamics.

[23] This is akin to autonomy in the SM case, although not literally as there are no differential equations.

Thus, as in the ZZW framework, there are two routes to predictions about later times: cf. Figure 3.2.

In both cases —ascending to the glider level of description from the cell level of description and ascending to a coarse-grained level of description (ρ_r) from the fine-grained description (ρ)— new and surprising features emerge.[24] In the Game of Life at the glider-level of description, there is 'motion'. At the cell-level there is no motion. Likewise in SM: at the coarse-grained higher-level of description, many features are different. The coarse-grained probability density ρ_r, the C^+ dynamics and the coarse-grained entropy S_{cg} are very different from their fine-grained counterparts: the fine-grained distribution ρ, the microdynamics, $U(t)$, and the fine-grained entropy, S_{fg}. In the paradigmatic case of N particles in a box, the two descriptions give different answers regarding whether the dynamics is reversible or not: in particular, about whether the Gibbs entropy increases over a period of time or not.

Admittedly, there are differences: in the SM case, there are no patterns that can be 'seen at a snapshot'. And because SM describes the evolution of probability densities there is no clear ontology at the higher-level description like Life's menagerie.[25] The pattern is the non-decreasing value of a particular quantity: the coarse-grained entropy, S_{cg}. This is not a synchronic pattern but a dynamical pattern. Furthermore, unlike the Game of Life case this is not a *visual* pattern. However, patterns at higher levels of description need not be "visual patterns but, one might say, *intellectual* patterns" that are "there for the picking up if only we are lucky or clever enough to hit on the right perspective" (Dennett, 1991, p. 41).

Yet, this in no way undermines its credentials as a pattern. One criterion for a higher-level pattern is predictive success: and betting that the coarse-grained entropy associated to an irreversible process will increase is a safe bet. Consequently, there 'are macroscopic patterns running through those very microscopic interactions' (O'Connor and Wong, 2015, 1.4) in both the SM and Game of Life cases.

To summarise: the important consequence of coarse-graining, i.e. of abstracting, is that autonomous dynamical patterns —structural features— once obscured by irrelevant details are revealed. Equipped with this alternative justification, I can now give a reply to the $(Illusory)$ objection in Section 3.6; and to the $(Anthropocentric)$ objection (in Section 3.7).

[24] Of course, in both cases, finding these features will depend sensitively on how the higher-level variables are defined, i.e. on how we abstract: cf. Knox (2016, p. 45).

[25] Note, also, that Life differs in another way: the patterns in Life are noise-intolerant — 'debris' can easily destroy the menagerie.

$$\begin{tikzcd}
\rho_r(t_0) & \rho_r(t_1) & & & \rho_r(t_n) \\
\rho(t_0) \arrow[u, "\hat{P}"'] \arrow[r, "U(t)"] & \rho(t_1) \arrow[u, "\hat{P}"'] \arrow[r, "U(t)"] & \rho(t_2)\dots & {} \arrow[r, "U(t)"] & \rho(t_n) \arrow[u, "\hat{P}"']
\end{tikzcd}$$

Figure 3.3: Route 1: to find the coarse-grained distribution ρ_r at any given time, evolve the full-distribution under the microdynamics $U(t)$ until this time and then apply the coarse-graining map $\hat{P}$.

$$\begin{tikzcd}
\rho_r(t_0) \arrow[r, "C^+(t)"] & \rho_r(t_1) \arrow[r, "C^+(t)"] & \dots & {} \arrow[r, "C^+(t)"] & \rho_r(t_n)
\end{tikzcd}$$

Figure 3.4: Route 2: to find the coarse-grained distribution ρ_r at any given time, evolve ρ_r under the C^+ dynamics until that time. Recall the C^+ dynamics is composed of applying U for Δt, applying $\hat{P}$, applying U for Δt.., where Δt is much smaller than $t_1 - t_0$.

3.6 Reply to $(Illusory)$

Recall that two premises were required to establish the conclusion that the asymmetry is illusory. According to the $(Illusory)$ objector's P1: coarse-graining distorts the correct density ρ. Furthermore, the coarse-grained asymmetry exists merely in virtue of the repeated coarse-graining every Δt in the C^+ dynamics (P2). Thus, as the asymmetry is rooted in the infidelity of coarse-graining, it is illusory.

The immediate reply to $(Illusory)$ is surely — the irreversible equations of SM are empirically adequate. If the asymmetry were illusory then we could not expect such success. Whilst this removes much of the force behind $(Illusory)$, the illusory objector might deny our assumption of empirical adequacy. In any case, in this Section I argue that P2 is false and this refutes $(Illusory)$. Furthermore, the considerations of Section 3.5 reveal that P1 is also false.

Contra to P2, the asymmetry is not generated *merely* in virtue of the continual coarse-graining — provided that the forwards-compatibility condition is met, the asymmetry is robust with respect to the number of applications of $\hat{P}$. Even if we eschew the C^+ dynamics, we could determine ρ_r at particular times t_1, t_n by evolving ρ under the microdynamics then projecting up to ρ_r at t_n. Call this route 1 (as shown in Figure 3.3, a version of the forwards-compatibility diagram in Section 3.3.1). Taking route 1, we would still find that the coarse-grained variables, ρ_r increase in entropy toward the future; $S(\rho_r(t_0)) \leqslant S(\rho_r(t_1)) \leqslant S(\rho_r(t_2))$. As such, we find an asymmetric pattern in ρ_r without using the C^+ dynamics. Thus, the asymmetry is not solely due to the repeated coarse-graining in the C^+ dynamics and so, P2 is false.

P1 claims that the action of $\hat{P}$ is to deliberately distort the correct density. That is, coarse-graining is a Galilean idealisation. On such a conception, ρ_r and ρ are analogous to the first term and full series respectively. According to $(Illusory)$, neglecting these higher-order terms is the source of the asymmetry. However, Section 3.5 revealed that coarse-graining is not a Galilean idealisation but rather an abstraction. ρ_r is not a distorted replacement but a new variable germane to a higher-level of description. Consequently, P1 is false.

Ultimately, however, the falsity of P2 is key to rebutting $(Illusory)$. The forwards-compatibility condition shows that the irrelevant degrees of freedom do not matter as they do not influence the evolution of the relevant degrees of freedom; they are not 'difference makers'. As such, the coarse-grained asymmetry would be robust — even if coarse-graining were a Galilean idealisation.

3.7 Reply to $(Anthropocentric)$

The anthropocentric objection is that no law determines the size of the cells and so we have a choice over which $\hat{P}$ to pick, and thus the coarse-grained quantities such as S_{cg} are anthropocentric. The concern was that this marks SM out as a theory worryingly different from the rest of physics.

However, my proposed alternative justification (Section 3.5.1) claims that the choice of coarse-graining map depends upon whether it uncovers successful autonomous dynamics, not our limited capacities. Thus it is not that we have a choice over which $\hat{P}$ to pick (and consequently the resulting equations and S_{cg} are 'tainted' by anthropocentrism). Rather it is a matter of whether ρ_r and ρ_{ir} dynamically decouple and "we are lucky or clever enough to hit on the perspective" — $\hat{P}$ — that reveals the patterns that are "there for the picking up" (Dennett, 1991, p. 41). There is no freedom in the choice that makes it depend upon our cognition (in a way that differs from the rest of the scientific enterprise). Only for particular choices of $\hat{P}$ is there an autonomous dynamics — the choice needs to be "just right" (Uffink, 2010, p. 195). And this situation is not special. Like countless moves in physics —in particular, countless definitions of good variables— the use is justified by its success: where here 'success' means that autonomous dynamics are found.

Consequently, coarse-grained features need not be anthropocentric in a way different from other physical quantities and so in this matter, SM has the same status as any other scientific theory. Hence, coarse-graining does not lead to a specific anthropocentrism, which one might have been concerned would render SM incompatible with scientific realism. (Of course, there is another potential source of subjectivity or anthropocentrism specific to SM: the use of probability. But as I stated in Chapter 2, this is beyond the scope of this thesis).

However, as discussed in Section 5, different levels of description are useful for different purposes and what is deemed useful may be relative to our human interests. Here our measuring capacities and imprecision are certainly relevant. Were we the size of a Maxwell demon and endowed with an ability to manipulate gas molecules, violations of the second law of thermodynamics might seem plausible. From their microscopic perspective, the second law might not seem like an obvious regularity in nature.

In addition, which patterns are uncovered might depend upon our limited human capacities — whether we can 'hit on the right perspective'. For instance, there may be regularities in the movement of the centre of mass of Cambridge's population, but our cognitive abilities may make us unable to pick up these patterns. Which variables we find useful depends on which variables we can access, i.e. measure and manipulate. Thus, our measuring capacities will clearly influence the construction and confirmation of our scientific theories. But — crucially — the details of our measuring limitations are not needed to *justify* coarse-graining in SM.

The above considerations highlight a potential *general* anthropocentrism: our scientific theories may be irrevocably entwined with our cognitive abilities and pragmatic interests. But this is not the return of the earlier (*Anthropocentric*) objection, which was *specific*: that the coarse-grained features are anthropocentric in a way that differs from the other putative physical quantities. The alternative justification shows that coarse-graining need not mark out SM as subjective and so different from other theories, but this conclusion is nonetheless compatible with scientific theories in general containing some element of anthropocentrism.

3.8 The wider landscape: concluding remarks

In Section 3.4, one of the concerns about coarse-graining was whether the coarse-grained asymmetry is 'real' or not. Recall that Davies claims that this was "a matter of philosophy"; and indeed, in Section 3.8.1 I explain why this is so: briefly, whether the asymmetry is real or not depends on one's views about inter-theoretic relations. Then in Section 3.8.2, I consider what my proposed justification reveals about the nature of irreversibility in SM.

3.8.1 Inter-theoretic relations

To some extent, the ZZW framework provides a case study in inter-theoretic relations; SM is a distinct, higher-level theory from either classical mechanics (CM) or quantum mechanics (QM). In the wider literature on inter-theoretic relations, one key issue is the nature of the connections between the different levels. For instance, biology and psychology could be disunified descriptions operating at different levels of generality: in addition to not being 'reducible-in-practice', they might not even be reducible-in-principle (Bedau and Humphreys, 2008, p. 215). That is, there may be disunity between the psychological and biological levels of description. Cartwright (1999), for example, advocates such a patchwork view of the scientific enterprise.

Different philosophical accounts of reduction make different requirements on the notion, and some are more stringent than others. (For instance, there is debate about whether any bridge laws invoked by the reduction must ensure the lower-level theory *explains* the higher-level theory). Here I will not return to the details of different accounts of reduction, since in chapter 1 I advocated 'reduction-as-construction'. This case study is clearly an application of that account in practice. After all, the ZZW framework allows us to *construct* the equations of one theory (SM) from another (CM or QM).

But there is a further issue concerning inter-theoretic relations: what attitude should one have to the higher-level entities — realism or instrumentalism? Hence, as Davies says, whether one believes the coarse-grained asymmetry is 'real' is a matter of philosophy: it depends on your prior philosophical convictions about higher-level entities in the special sciences. But my conclusion in chapter 1 was that reduction vindicates the higher-level entities, and so I claim the higher-level asymmetry is real.

Furthermore, such philosophical convictions may also have a general impact on one's views about the nature of the asymmetry. Had the MI justification been the best justification of coarse-graining, then the coarse-grained asymmetry would have been revealed to be inescapably subjective or anthropocentric. Whilst I hope to have established (in Sections 3.6 and 3.7) that one is not *compelled* to consider the asymmetry to be anthropocentric, motivated by general themes in inter-theoretic relations: one might still want to conclude that it is, in fact, anthropocentric. For example, an instrumentalist about higher-level theories might maintain that the instrumental value of these descriptions is inextricably bound up with our measuring and cognitive capacities and thus, all higher-level entities are anthropocentric. The key message of this chapter is that the justification of coarse-graining need not mark SM out from other scientific theories as regards that general debate (as we saw in Section 3.7).

Next, there is a final philosophical issue about the nature of the coarse-grained asymmetry to discuss: its emergent nature.

3.8.2 The nature of irreversibility

Finally, I turn to irreversibility. As a foil for this discussion, I choose a passage from Sklar (1993), which puts very well a general doubt: whether a strategy such as the one outlined in this chapter, can really succeed in reconciling the time-symmetry of micro-processes with the asymmetry of macro-processes.

> "Do the procedures for deriving kinetic equations and the approach to equilibrium really generate *fundamentally* time-asymmetric results?" (Sklar, 1993, p. 217, emphasis added).

However, contra to Sklar's phrasing, the ZZW construction method does not generate a *fundamental* time-asymmetry. The coarse-grained asymmetry is a feature of a higher-level description. Higher-level descriptions can have features that differ substantially from the lower-level descriptions (without there being a contradiction). Often these features are described as *emergent*.

Agreed, 'emergence' is a murky word and is used in many different ways (see Silberstein (2002) for a survey). Very roughly, emergent entities or processes 'arise' out of more fundamental entities or processes and yet have 'distinctive' features in their own right. It is contentious what the 'distinctive' features are: proposals in the literature include 'novelty' (Butterfield, 2011b, p. 1065) and being 'unexpected' (Chalmers, 2006, p. 244).[26] Furthermore, how substantively a phrase such as 'in their own right' must be read also varies across authors — some maintain that emergence is the failure of reduction whilst others (e.g. Butterfield (2011b)) deny this. The menagerie of the Game of Life, such as gliders and blinkers, are often cited as key examples of emergent entities that have certain emergent properties and evolve under certain emergent processes (Bedau and Humphreys, 2008).

The sense in which I use 'emergent' is mild; it is merely that there is 'novel and robust behaviour with respect to some comparison class' (Butterfield, 2011b, p. 1065). (Butterfield's account is especially apt for this case, since he shows his definition to be compatible with inter-theoretic reduction, and as discussed above, the ZZW construction is a case of reduction).

Of course, as mentioned above, there are many accounts of emergence that one could favour. An alternative account that might seem apt here is Wilson (2009). Her key idea is that some phenomena are "weakly ontologically emergent from physical phenomena" (Wilson, 2009, p. 280) when some degrees of freedom are *eliminated*. Note that eliminating functional dependence of one set of degrees of freedom from another was exactly the autonomy condition of the ZZW framework. Furthermore, her accounts fits well with the general topic of abstraction and talk of levels of description. However,

[26] Both of these examples are definitions of 'weak emergence' (a use of the word 'emergence' popular with scientists and philosophers of physics) as opposed to the philosopher's 'strong emergence', which is a logically stronger notion. Although authors vary about exactly what the distinction between weak and strong emergence is, the idea is that this stronger sense implies a lack of reduction or supervenience of the emergent phenomenon on the lower level. See Chalmers (2006) for more detail on the weak/strong distinction.

Wilson's focus is on weakly emergent *entities* and as mentioned at the end of Section 3.5.2, one of the disanalogies with the Game of life is that is unclear in our case what the candidate emergent entities would be. Thus, I will not pursue Wilson's account further here. Instead: I submit that the broad gist of Butterfield's account captures the main intuition common to all accounts of 'emergent phenomena': *robust*, because a putative case of emergence must not be too flimsy in order to count as a bona fide phenomenon and *novel*, in order to earn the name 'emergent'.[27]

Thus, my response to Sklar's concerns above is as follows: the irreversibility generated by these methods is not fundamental but emergent. Irreversibility emerges when one abstracts from the fine-grained level of description to the coarse-grained level of description by applying a $\hat{P}$ that leads to autonomous dynamics.

Note finally that this mild conclusion that the coarse-grained asymmetry is weakly emergent is not "toothless". It is in direct opposition to Prigogine and Stengers (1984, p. 285) who claim: "Irreversibility is either true on all levels or on none: it cannot emerge as if out of nothing, on going from one level to another". Whilst the lower-level dynamics is reversible, the coarse-grained dynamics at the higher level of description is irreversible. True, this emergent irreversibility does not arise "as if out of nothing". Time-asymmetric assumptions were required when constructing the C^+ dynamics (and when ruling out the C^- dynamics) in Section 3.2. But this is to be expected; if no asymmetry is put in, then we cannot expect asymmetry out.

Since the time-asymmetry is not fundamental and was "put in by hand" (as discussed earlier), this project won't satisfy those seeking to locate the source of time-asymmetry. To the extent that this project answers that question, it claims the asymmetry arises because of particular initial conditions (the initial state assumption). Some want an explanation of such initial conditions, especially when in the guise of a 'Past Hypothesis' (cf. Callender (2004); Price (2004)), especially since such initial constraints seem ad hoc or unnatural from "the mechanical world-view" (Sklar, 1993, p. 368). Moreover, there is a debate over whether such an initial state is a law or a 'de facto' condition (Reichenbach, 1991; Grünbaum, 1973; Sklar, 1993, p. 370). Some such as Krylov (1979), are unhappy with the centrality of such initial conditions in explaining irreversibility.

But I believe my conclusion about the emergent nature of the asymmetry helps us to see which explanatory projects are likely to be fruitful. In particular, my conclusion eases the worry that the initial conditions required do not look especially natural nor form a unified class. Because these higher-level patterns are weakly emergent, they are *unexpected* from the lower-level mechanical perspective. Thus, the moves required at the lower-level in constructing SM equations may often look unnatural: otherwise

[27] As an aside: Wilson's aim in her account of emergence is defend non-reductive physicalism — I leave it to future work to consider how non-reductive physicalism relates to the non-eliminative reductionist picture painted in this thesis.

the higher-level pattern would have been expected. (This point was also discussed in chapter 1: the lower-level resources might not suffice for determining which states form an equivalence class, not least because there are so many possible abstractions, or coarse-grainings. Hence, the theoretical moves required to construct the higher-level theory from the lower-level theory might be surprising).

To sum up: the ZZW framework constructs the irreversible equations of SM from the underlying reversible microdynamics: thus, reconciling the higher-level asymmetry with the lower-level symmetry. The procedure of coarse-graining — key to this reconciliation but thought to be suspicious by many — was justified provided that coarse-graining allows us to abstract to a higher-level autonomous description (in a manner illustrated by the Game of Life). I used my justification of coarse-graining to show that the coarse-grained asymmetry is neither illusory nor anthropocentric, but instead: weakly emergent.

4 Stars and steam engines: to what extent do thermodynamics and statistical mechanics apply to self-gravitating systems?

4.1 Introduction

The foundations of thermal physics are riddled with controversy. The keystone of the philosophical debate is the old issue of whether thermodynamics (TD) reduces in some appropriate way to statistical mechanics (SM). This issue involves analysing the thermodynamic limit (i.e. roughly the limit of an infinite number of microconstituents); and so leads immediately to the topic of infinite idealisations. Namely: how should we understand this limit given that any physical system to which we successfully apply thermodynamics and/or statistical mechanics contains in fact a finite number of atoms (or other microconstituents)? The debate in thermal physics has centred around phase transitions. The SM description of phase transitions requires the thermodynamic limit, unlike the TD description, which seemingly makes the TD description superior and so—some argue—non-reducible to SM. In this chapter, I consider a different case in thermal physics where, I will argue, statistical mechanics has the upper hand. This case further differs from the usual phase transitions case: I will argue that the philosophical interest of this field of physics, and the light it sheds on the thermodynamics/statistical mechanics relation, turns on the fact that here, the thermodynamic limit does *not* exist.

This field is often called 'gravitational thermal physics' — and so the tangles of thermal physics reach beyond our terrestrial sphere. But even if we set aside black holes, the claim that thermal physics successfully applies to *Newtonian* astrophysical contexts has been disputed. Such an enterprise involves applying the ideas of thermal physics to vast collections of stars: both globular clusters with ca. 10^5 stars and galaxies with ca. 10^{11} stars. The key idea is to think of such a collection as like a gas: just as the molecules in a gas are its microconstituents, the stars in such a collection are *its* "microconstituents". This is obviously a very striking, indeed bold, idea: both physically and philosophically. Physically, because we expect disanalogies between the idealisations made for a collection of molecules and those made for a collection of stars. In particular, stars interact by gravity, which is systematically set aside in terrestrial applications of thermal physics. Philosophically, because our epistemic access to (our warrant for believing in) molecules and stars are so very different. Stars are epitomes of the observable; since the ancients turned their eyes heavenwards, we have believed in them — though of course what we have believed *about* them has altered immensely since ancient times, especially since 1850 with the application of spectroscopy to starlight through to today's stunning observational knowledge of stars' lifecycles. This philosophical disanalogy between molecules and stars will play out in what follows, especially in connection with (1) the relationship of thermodynamics to statistical mechanics and (2) Einstein's distinction between constitutive and principle theories. And as we will see, the question of the existence and the nature of the thermodynamic limit — the infinite idealisation of infinitely many stars — will be central.

Whilst such philosophical and physical disanalogies abound, ultimately the question is whether (Newtonian) gravitational thermal physics is a successful enterprise. Thus Callender asks whether "the stars in such systems or even the galaxies themselves, when idealised as point particles, admit a thermodynamic description" (Callender, 2010, p.44). Does thermal physics apply to these Newtonian self-gravitating systems? Is this an extension of the domain of applicability? Indeed, does this case give further weight to the idea that thermodynamics is universal?

On the one hand, it seems that thermal physics applies to self-gravitating systems (SGS). For instance, in certain circumstances the evolution of the distribution of stars in a galaxy can be modeled using the collisionless Boltzmann equation or the Fokker-Planck equation (see e.g. Binney and Tremaine (1987)). On the other hand, self-gravitating systems exhibit many unusual features, sometimes called the 'gravitational paradoxes', as discussed in section 4.2 .

There is a prima facie dispute in the scientific community. Some express Optimism over the applicability of thermal physics: "Statistical mechanics of gravitating systems is a controversial subject. However, our modern understanding of statistical mechanics and thermodynamics does handle gravitational interactions rigorously with complete satisfaction" (Kiessling, 1999, p. 545). Other express Pessimism: "[Thermodynamics] is essentially a human science; it started with steam engines and went on to describe many physical and chemical systems whose size is of the order of a metre. They clearly are inapplicable to the solar system or to galaxies. Clearly classical thermodynamics is not a useful branch of science in cosmology; we have extrapolated too far from its human-sized origins" (Rowlinson et al., 1993, p. 873).

Of course, one might be tempted to 'hedge your bets' and claim that whilst gravitational thermal physics has some successes, this success is qualified by the paradoxes. That is, one might claim, as is often the case with optimism and pessimism, that there are shades of grey: the truth lies in between.

But I think we can do better than merely hedging our bets in this dispute. My goal in this chapter is to make peace between the Optimists and the Pessimists, by deflating the debate between them. I argue that: if we are careful in distinguishing statistical mechanics and thermodynamics, then no reconciliation is required. Both sides can live in harmony because whilst statistical mechanics applies, thermodynamics does not.

This position differs from Callender (2011), who brought this dispute to the attention of the philosophy of physics community (Callender, 2010). He notes the successful features emphasised by the Optimists, whilst not minimising the difficult features that a Pessimist might stress. But — motivated by his broader position in the foundations of thermal physics, namely: we should not take thermodynamics too seriously and so advocates a more flexible, and so more liberal, view of thermodynamics — Callender subscribes to a (cautious) Optimism. That is, he holds that the problems facing SGS do not "spell the end for gravitational equilibrium thermodynamics" (Callender, 2011, p.962).

A disclaimer at the outset: my reply to Callender will not hinge on bringing new physics to bear on the dispute, but rather a different perspective on the foundations of thermal physics. Thus the main message will be: the example of self-gravitating systems need not necessitate having a broader or more flexible view of thermodynamics.

In section 4.2 , I recapitulate Callender's discussion of the thermal physics of self-gravitating systems: the difference from ordinary systems, the successes and the unusual 'gravitational paradoxes'. Section 4.3 outlines my strategy of delineating SM and TD, and connects such a strategy to the reduction debate, and to Callender's position. In section 4.4, I argue that thermodynamics does not apply to self-gravitating systems. Section 4.5 outlines the extent to which statistical mechanics applies. Thus, my verdict on the dispute is that there is (to an extent) a statistical mechanical description of SGS, even though no thermodynamic behaviour emerges. Section 4.6 sketches an explanation of why thermodynamics and statistical mechanics come apart in this case. This explanation will hinge on the thermodynamic limit, and so I also outline the connections to the wider debate about the role of the thermodynamic limit in the relationship between SM and TD. Section 4.7 concludes.

4.2 Newtonian Gravity weighs in

In this section, I first review how incorporating gravity changes the physics of thermal systems and discuss the type of systems well-approximated by this treatment. I then outline two examples of successful evolution equations. Finally, I discuss some of the 'paradoxical' features.

4.2.1 How the situation changes with gravity

Gravitational forces are negligible in the terrestrial thermal systems with which we are familiar. But in extraterrestrial systems such as galaxies, this assumption is of course no longer justified. Unlike the local collisions and forces in an ideal gas, the gravitational force is *long-range*; the range of the dominant interaction is large relative to the spatial size of the system. Consequently, the forces on a given star are not only due to its nearest neighbours, but include a contribution from the large scale structure of the stellar system. Indeed, if the density of stars is spatially constant (cf. Figure 4.1), the gravitational force exerted on a given star (by the rest of the stellar system) at the apex is the same from the patch of stars of solid angle $d\Omega$ surrounding it at distance r_1, as from the patch at distance r_2. Clearly if the distribution of stars were exactly spherical, there would be no net force on this star. However, if the density of stars falls off more slowly in one direction, then only this very global feature of the entire stellar system will be responsible for the force on our star. This contrasts sharply with the forces experienced by a molecule in gas, which come only from its nearest neighbours and thus is a much more local feature.

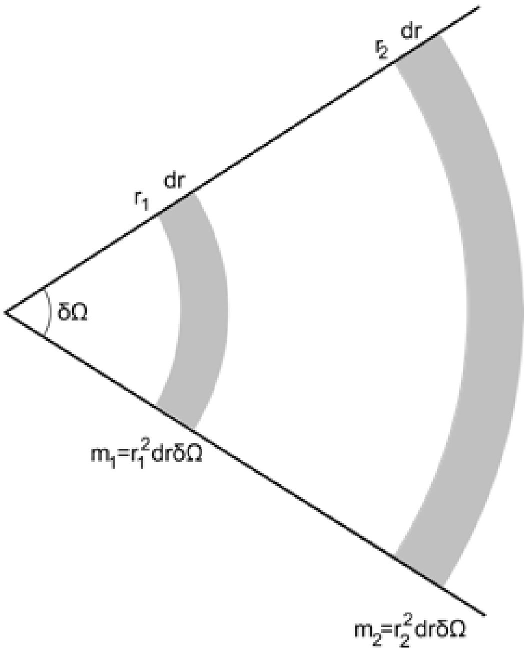

Figure 4.1: Diagram from Callender (2010) (adapted from Binney and Tremaine). In a collisionless system with constant density of stars, the force exerted on a star at the apex is the same from the band of stars r_1 as from the band of stars at r_2.

The gravitational potential $V \sim \frac{1}{r}$ is asymptotically zero; and this dominates the behaviour of SGS due to (i) its infinite range and (ii) the fact that a (potentially infinite) amount of energy can be released as two point particles get arbitrarily close together, as seen in Figure 4.2.

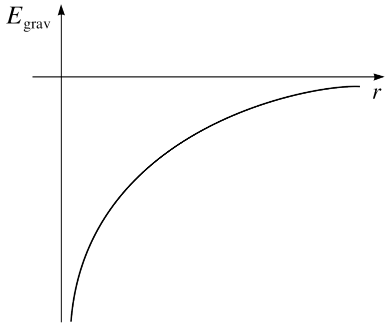

Figure 4.2: The gravitational potential energy. Here r corresponds to $|q_i - q_j|$ in equation 4.1.

Here, we primarily focus on the gravitational n-body case where stellar systems are treated as collections of n point masses. Unlike ideal gases, the total energy is not even approximately the sum of the kinetic energy of the constituents since the (negative) gravitational potential energy must be included. The Hamiltonian for such a system of n 'particles' of equal mass m is thus:

$$H(\mathbf{q}, \mathbf{p}) = \sum_{i=1}^{n} \frac{\mathbf{p_i}^2}{2m} - \frac{1}{2} \sum_{i=1}^{n} \sum_{i \neq j} \frac{Gm^2}{|\mathbf{q_i} - \mathbf{q_j}|} \tag{4.1}$$

Whilst this is an idealisation, it provides a very successful description of elliptical galaxies (10^{11} stars) and globular clusters, i.e. spherical gravitationally bound systems of about 10^5 stars, which both contain very little interstellar medium (dust and gas).

Of course, for some systems we cannot ignore hydrodynamics — namely when interstellar dust and gas are relevant. And for some systems general relativity cannot be ignored. For example, this applies when black holes are present, and when the cosmological structure i.e. curvature of space on very long length scales, cannot be ignored, such as in the dynamics of clusters of galaxies.

Indeed, I should make an obvious and more general disclaimer: whilst Newtonian thermal physics can be used in galactic dynamics describing extraterrestrial systems it is (unsurprisingly) far from the whole story. Nevertheless, models based on the simple Hamiltonian (4.1) have had some venerable successes: cf. §4.2.2.

4.2.2 Successes

I shall sketch two approaches, the first assuming stars do not 'collide', the second allowing for collisions. To model these gravitating systems, the broad idea is to find a probability density function f in phase space and consider its evolution.

Modelling a stellar system to be collisionless requires the approximation that no 'encounters' occur. An encounter occurs when two stars are so close as to cause a gravitational perturbation, altering their orbits. (Collisions involving physical contact between stars are exceedingly rare and can be ignored in most models.)

The star's orbit is then approximated by assuming the total mass of the system is smoothly distributed instead of concentrated in point-like stars. This 'collisionless' (encounter-less) approximation holds for certain systems, in particular: for globular clusters and elliptical galaxies (containing about 10^{10} stars) since, for timescales less than the relaxation time, stellar encounters are unimportant except at their centres (Binney and Tremaine, 1987).

Here, the relaxation time is proportional to the number of stars and the time taken for a star to cross the galaxy (the crossing time). After the relaxation time the star's actual velocity differs from the smooth gravitational field case and its orbit will deviate from the smooth field model by an amount of the order of its original velocity.

As in Boltzmann's treatment of a dilute gas, we define a probability density function $f(\overrightarrow{r}, \overrightarrow{v}, t)$ where $f(\overrightarrow{r}, \overrightarrow{v}, t)d^3rd^3v$ gives the probability at t of finding a star in volume d^3r around r with velocity within d^3v of v. Since we assume all N stars have the same probability density function (and are stochastically independent of each other), this function is defined in a 6-dimensional phase space, rather than the $6N$-dimensional phase space of the entire set of N stars.

The collisionless Boltzmann equation gives this function's evolution;

$$\frac{\partial f}{\partial t} + [f, H] = 0 \tag{4.2}$$

where H is given by equation 4.1. Note that the collisionless Boltzmann equation is nonlinear as the gravitational potential $\Phi(x, t)$ depends on the distribution of stars' masses, $f(\overrightarrow{r}, \overrightarrow{v}, t)$.

We can define the entropy

$$S = -N \int f(\vec{r}, \vec{v}, t) \ln f(\vec{r}, \vec{v}, t) d^3rd^3v. \tag{4.3}$$

To look at the evolution of a stellar system over timescales longer than the relaxation time, in which encounters between stars must be considered, we need what is (usually) called the Fokker-Planck approximation. The encounter operator, $\Lambda[f]$, gives the difference of the probability that a star is scattered into and out of a volume of phase space in a given time interval. Equation 4.2 becomes

$$\frac{\partial f}{\partial t} + [f, H] = \Lambda[f]. \tag{4.4}$$

To sum up: the collisionless Boltzmann and Fokker-Planck equations have proven to be empirically successful evolution equations for the systems described at the end of §4.2.1.[1]

4.2.3 Unusual features

However, the extension of thermal physics to SGS is far from seamless. There are a wide array of problems surveyed in Callender (2011): of which I will consider only three.

(1) *Strong interactions*. Firstly, functions, such as energy and entropy, are often not additive or extensive for SGS. For an ideal gas the total energy E is the kinetic energy K, whereas for gravitating systems the (negative) potential energy U contributes: $E = K + U$. Functions such as energy and entropy are usually additive: the energy of a combined system A+B is just the sum of the energy of A and the energy of B. Usually the Hamiltonian of the joint system is $H_{AB} = H_A + H_B + H_{int}$, but it can be approximated by $H_{AB} = H_A + H_B$. (So strictly speaking, the energy is additive iff there are no interactions, i.e. $H_{int} = 0$). However, a SGS will not have even *approximately* additive functions since the neighbouring stars do not contribute the majority of the influence on a particular star (Cf. Figure 4.1). That is, the interaction Hamiltonian, $H_{int} \not\approx 0$. The physical reason for this can be seen in Figure 4.3 and 4.4, showing how putting together two 'boxes' of gravitating stars alters both boxes: the long-range attractive forces result in 'clustering' or 'clumping' not seen for ideal gases (or indeed real gases in terrestrial settings, which are well described by zero or only short-range forces between constituents). For these gases, short-range potentials are dominant — adding two boxes of gases does not alter the systems in such a dramatic way, since the systems only interact at their boundary.

As a consequence, variables such as energy and entropy are usually taken to be *extensive*. Here, a variable is called 'extensive' if it depends linearly on the size of, i.e. the number of constituents in, the system (e.g. mass, internal energy, volume)[2] and is called 'intensive' if independent of system size (e.g. density, pressure). The energy of a subsystem is proportional to the volume, whereas interactions between subsystems are proportional to their interface boundary's surface area and are, therefore, of a smaller order of magnitude, provided the subsystems are big enough. So strictly speaking, even for short-range potentials, entropy and energy are only extensive in the thermodynamic

[1] For some examples of solutions to the collisionless Boltzmann equation, the initial conditions and approximations involved, see (Heggie and Hut, 2003, ch. 8). Much of the research in this area focuses computational simulations. One useful class of models that take the velocity distribution to be isotropic is Plummer's model, named after Plummer who used this approximation to fit the observed light distributions of clusters (Spitzer, 1987, p. 13).

[2] More generally, an extensive variable Q is homogeneous (in the modes n_c) iff $Q(kn_1, kn_2, ..) = kQ(n_1, n_2, ...)$.

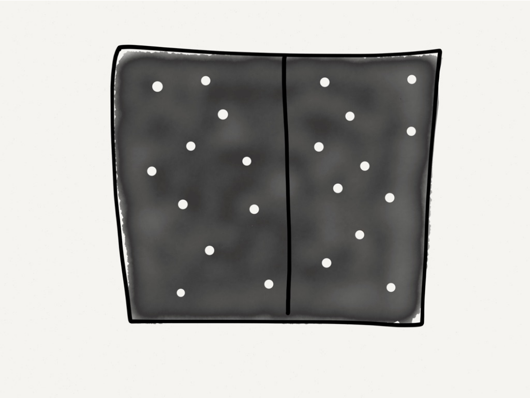

Figure 4.3: Gas molecules in a box.

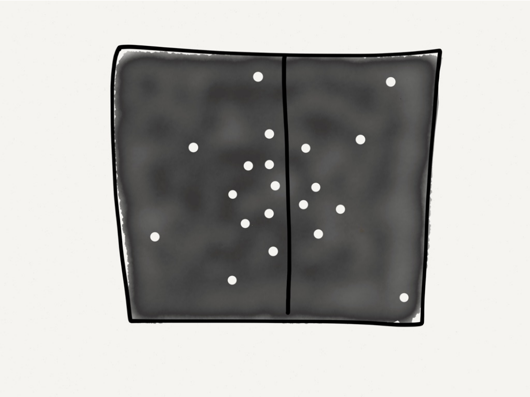

Figure 4.4: Stars in a box.

limit. But although this is a matter of degree, there is still a contrast of principle with SGS. For energy and entropy are not extensive for gravitating systems, no matter how large the system.[3]

(2) *Putting in energy reduces the temperature.* Gravitating systems can have a very unusual property: negative heat capacity. The *heat capacity* (at constant volume) is the amount of energy required to raise the temperature by one degree at constant volume;

$$C_V = \left.\frac{\partial E}{\partial T}\right|_V . \tag{4.5}$$

When the system is in virial equilibrium (where $2K + U = 0$), the total energy is negative ($E = K + U$, so $E = -K$, where K is by definition positive). From the equipartition theorem, we have $K = \frac{3}{2}Nk_BT$. This implies $E = -\frac{3}{2}Nk_BT$ and thus

[3] Callender (2011, p.974) takes the failure of extensivity to be a key problem, but suggests that perhaps extensivity can be recovered through the Kac prescription (a rescaling of the temporal and spatial parameters such that the constant $c := -Gm^2$ in the Hamiltonian (1) becomes $c = \pm\frac{1}{N}$). Since the merits of this prescription is an open issue in physics, I do not explore it further here.

$C_V = -\frac{3}{2}Nk_B$: the heat capacity is negative. If the system gives out energy, the temperature will increase. If you put energy into a system, the temperature goes down. Indeed, unusual!

(3) *The gravothermal catastrophe*: Thirdly, there is the infamous gravothermal catastrophe (Lynden-Bell et al., 1968). To explain this, let us consider in general terms which evolutions are entropically favourable. Whether a process (such as expansion) increases entropy depends on whether the phase space volume increases. Thus, for example, expansion of an ideal gas is entropically favoured since it increases the volume available. Ceteris paribus, the hotter the system the higher its entropy as more momentum states are available (due to the increased kinetic energy). So whether an expansion of a self-gravitating system increases or decreases entropy depends on how the competing factors affect the phase space volume (Wallace, 2010). An increased volume means more spatial states but results in a decreased number of momentum states as the kinetic energy has decreased, since work is done against the attractive gravitational field.

Turning now to SGS: when the density contrast between the edge and centre of a SGS is great enough, we conceptually divide the system into a uniform core and a uniform halo, each in virial equilibrium. If a small amount of heat is transferred to the envelope from the core, the core's kinetic energy decreases, making it favourable for the core to contract (as $U = 2E$, E has decreased so U is more negative). Since the core has negative heat capacity, losing energy increases the temperature. The core decreases in entropy but this is more than offset by the expansion and cooling of the halo.[4] The heat flow and contraction increases the temperature gradient between the core and envelope and thus the process of heat transfer from the core to the halo is self-perpetuating.

The gravitational potential, $V \sim \frac{1}{r}$, being unbounded from below as $r \to 0$, means that this collapse would appear to continue without end. For an infinite amount of potential energy can be released by moving two particles closer and closer together, as seen in Figure 4.2. Consequently, it seems that there are no equilibrium states. No equilibrium will be reached since, according to the gravitational potential, the core can keep contracting indefinitely becoming infinitely dense.

Is this gravothermal collapse observed? Here we meet a familiar philosophical theme: that singularities in one theory can signify the breakdown of that theory, and often signal some features of the successor theory (Berry, 2002; Batterman, 2001) — so that idealisations taking some quantity to infinity can play a key role in inter-theory relations. More generally, physics consists of models which have a limited domain of applicability; if you push any model of physics far enough it will break down. As Feynman quips: "When you follow any of our physics too far, you find it always gets into some kind of trouble" (Feynman et al., 1964, §28.1). The same point is made

[4] Conservation of energy requires that the heat flow from the core to the halo increases the halo's energy — which is now *less* negative. Thus, it is favourable for the halo to expand and cool.

in the literature about SGS: Hut says "whenever a theory predicts the occurrence of singularities, it has been a sign that other physical effects, which have been overlooked, will kick in before actual infinities are reached" (Hut, 1997).

But to return the question of gravothermal collapse: indeed, as Hut says, other physical effects eventually kick in. Globular clusters undergo this gravothermal collapse, albeit over a period of tens of millions of years. Agreed: in a globular cluster, the formation of hard binaries provides the core with an energy source (Spitzer and Ostriker, 1997, p. 363): nevertheless, once exhausted gravitational collapse will continue. Another instance is a contracting gas cloud (that ultimately will form stars) where the heat is emitted as electromagnetic radiation (due to the presence of an interstellar medium which is absent from globular clusters). In the case of stars, fusion processes provide the energy source to resist gravitational collapse but eventually this energy source runs out. In this case, gravitational collapse resumes until another effect (dependent on the star's mass) kicks in. For example: for stars of around 10 solar masses, collapse continues until a supernova occurs leaving a neutron star in which the degeneracy pressure (a consequence of the Pauli exclusion principle) resists the attractive force of gravity (Phillips, 2013).

But I will not need more details about these "additional physical effects". For this chapter, the main point of all these other effects is that they involve various theories and subdisciplines of physics such as hydrodynamics, quantum theory — and statistical mechanics.[5]

4.3 My Strategy for Reconciliation

Callender (2011) argues that to reconcile the two sides of the debate, we should take a broader, more liberal view of thermodynamics. We should not 'take thermodynamics too seriously' but allow for such unusual features. For example, equilibrium needn't be strict (Callender, 2001).

Thus Callender asks: what should we conclude from SGS's unusual features? He says *"If there is a general lesson, I believe it is that we sometimes have too narrow a vision of thermodynamics. In his beautiful review, Thompson (1972) writes that 'to show that thermodynamics exists for a given system' we must (a) 'prove. . . the existence of the thermodynamic*

[5] Callender attempts to abstract away from these details above by altering the gravitational potential by introducing a short cut-off potential η which prevents the gravitational potential $\rightarrow -\infty$ as $|q_i - q_j| \rightarrow 0$ as the potential is now bounded from below.

$$\frac{1}{|\mathbf{q_i} - \mathbf{q_j}| \rightarrow \frac{1}{\sqrt{(\mathbf{q_i} - \mathbf{q_j})^2 + \eta^2}}} \tag{4.6}$$

Perhaps this short distance cut-off is artificial, but regardless of whether we impose it, there are still no equilibrium states in the sense of a state with maximum entropy.

limit' and (b) 'show that the resulting thermodynamics is stable', i.e., prove that specific heat is positive. By these criteria, self-gravitating systems badly fail as thermodynamic systems. Yet thermodynamic techniques sometimes have proven successful when applied to self-gravitating systems. How do we reconcile these two facts?" (Callender, 2011, p. 979).

I advocate a different view: by dividing thermal physics into thermodynamics and statistical mechanics, no reconciliation is required. This is because phenomenological thermodynamics does not apply to these systems (a claim I argue for in Section 4.4), although, to a certain extent, statistical mechanics does (a claim I argue for in Section 4.5). Thus, the dispute over the applicability of thermal physics is deflated as merely semantic: the Optimists are talking about SM whereas the Pessimists are talking about TD.

Of course, dividing thermal physics into TD and SM is an incredibly contentious matter. Can one draw a clean line and if so, where should one draw it? I submit that some division, albeit a rough or vague one, must be possible, as a prerequisite of the meaningfulness of the reduction debate, which after all requires that there are *two* theories, one of which may or may not 'reduce' to the other. (And whatever one's qualms about the reduction debate, to say it is meaningless is surely just intellectual defeatism).

That such a line can be drawn is a prerequisite of the reduction debate and *how* it is drawn is important for whether a reduction exists: for claims of reduction are evaluated not only in relation to a given account of reduction, but also in relation to the definitions of the two theories.

I agree that there are multiple possibilities of how to draw the line between TD and SM — and these different options have various foundational motivations. There is a plurality of ways to carve up the terrain, and how one does it depends on one's aims. Thus if you believe that SM is the powerhouse of thermal physics (Wallace, 2015a), your preferred line might be different from those who venerate thermodynamics (such as (Eddington, 1928, p. 104)). In addition to this question of the conceptual priority of TD or SM, the foundational debate between Gibbsians and Boltzmannians plays a role. For instance, one might advocate a Gibbsian definition of equilibrium (that the probability distribution is stationary) because it nicely lines up with the thermodynamic definition (that the macrovariables are stationary). This is because the phase average of a macrovariable[6], using a stationary probability distribution will also be stationary - so with reduction in mind, this Gibbsian definition is a good SM candidate for reducing TD equilibrium. On the other hand, Callender advocates a Boltzmannian view of SM: according to which equilibrium is defined to be the largest macrostate and the system can fluctuate away from equilibrium, which is arguably unlike the traditional

[6] Of course, why Gibbs phase averaging works is a source of controversy, cf., e.g. Malament and Zabell (1980).

thermodynamic definition. In order that reduction is still on the cards, Callender advocates taking a more liberal view of thermodynamics (Callender, 2001). Thus, Callender's reconciliation between the Optimists and the Pessimists is part of his wider view of the foundation of thermal physics. Hence, the line I propose between TD and SM in what follows may have different foundational motivations than that of Callender and others.

Thus, whilst some split must be possible, that is not to say it is either precise or wholly objective. Thermodynamics and statistical mechanics developed 'cheek by jowl' and so a sharing, indeed a blurring of concepts, methods and results seems inevitable. As part of the historical progress of science, the original meaning of theoretical terms in one theory bends under the success of another. As is well-documented, the success of SM led to conceptual extensions of TD, such as negative temperatures. Indeed, one might see this as evidence of a successful horizontal reduction, with SM the successor theory. Furthermore, this explains how such a semantic dispute could arise between the Optimists and the Pessimists; often physicists talk of 'statistical thermodynamics' and arguably it is the (putative) reduction that has led to the blurring of the two theories for practical purposes.

But the putative reduction of TD to SM is not only a diachronic reduction of an older theory to its successor, but also a synchronic reduction of the higher-level macroscopic theory to the lower-level underlying microscopic theory. The inter-theoretic relationship between TD-SM differs from the classic horizontal reduction between Newtonian mechanics and Special Relativity. Newtonian mechanics is wrong in certain domains and predictions, but it is contested whether thermodynamics is wrong in the same way. This is made especially clear by those who venerate thermodynamics, claiming it to be fundamental. Planck, for example, took the second law of TD to be universal, applicable to "every process occurring in nature" (Planck, 1926, p. 463) (as quoted in (Uffink, 2006b, p. 280)). One classic exponent of this view is Eddington, who claimed that: *"If someone points out to you that your pet theory of the universe is in disagreement with Maxwell's equations - then so much the worse for Maxwell's equations. If it is found to be contradicted by observation - well, these experimentalists do bungle things sometimes. But if your theory is found to be against the second law of thermodynamics I can give you no hope; there is nothing for it but to collapse in deepest humiliation"* (Eddington, 1928, p. 104). Such a view is certainly at odds with comparing TD to the (superseded) Newtonian theory when discussing horizontal reduction. Thermodynamics is not straightforwardly 'inferior' to statistical mechanics.

But in the case of SGS I will argue: TD does not apply but SM does. And it is of foundational significance *which* theory these successes of thermal physics in this exotic domain belong to. This is because of the above question of the conceptual priority of TD and SM, which influences the reduction debate. For instance, one possible—if not popular—position is that it is a failing of the Boltzmannian entropy that it does not *strictly* increase, whereas it is a virtue of the Gibbs coarse-grained entropy that it does: because this is more faithful to the thermodynamic entropy (cf. (Callender, 2001)). Thus, not only must we have two distinct theories and a definition of each, but we also need to be clear on the conceptual priority of one over the other.

Despite these connections and complications surrounding reduction and the sharing, indeed blurring of concepts, I contend that two core frameworks can be distinguished — though of course, boundary cases may still remain. In broad terms, this goes as follows. TD is an abstract theory, that proceeds in ignorance of the constitution of the system, dealing instead only with macrovariables which obey the Four Laws (or really, Five Laws — cf. Section 4.4 on the "minus first law"). In contrast, SM describes systems by considering statistical, or probabilistic, distributional features of the microvariables. In particular, the state space of equilibrium thermodynamics consists of equilibrium states labelled by a small number of macrovariables, whereas the state space of statistical mechanics consists of appropriate probability density distributions over microvariables, such as position and velocity of the microconstituents. In order that we do not beg the question about reduction, it is important that we keep the concepts of each theory distinct. Accordingly, the concepts of SM, in particular a SM notion of entropy or equilibrium *may* turn out to identical to the thermodynamic entropy or equilibrium — but this would be a major case of theoretical identification and so should not be assumed at the outset.

Having admitted the difficulties with dividing thermal physics into SM and TD, I now offer two reasons why my position — the debate between the Optimists and Pessimists can be deflated, because TD does not apply although (to an extent) SM does — might be anticipated/seem natural.

Firstly: as highlighted in the introduction, TD was created in a time when there was much scepticism about the existence of atoms. Because of this uncertainty surrounding atoms, TD arose as a theory of empirical generalisations about the bulk properties of matter, without regard to its microscopic composition. Einstein famously called thermodynamics a 'principle theory', in contrast to those 'constructive theories' that 'build complex phenomena out of relatively simple postulates' (Einstein, 1919, p.228).[7] The generalisations of TD were extrapolated from regularities in phenomena familiar from tabletop systems of gases, pistons etc. Thus it is unsurprising that these generalisations do not hold in the radically different realm of stars and galaxies. But the constructive theory now considered to underpin the generalisations of TD—SM—may well apply (and this is considered in Section 4.5).

[7] This distinction, though announced in a 'mere' newspaper article has had a great legacy, e.g. in the debate about the primacy of matter vs. geometry in the philosophy of spacetime, e.g. Brown (2005); Janssen (2009).

Secondly: my view is already suggested by some of the physics material reviewed above. The thermal physics of SGS never abstracts away to macroscopic bulk variables from the microvariables —i.e. the position and momenta of the individual stars— and probability distributions over these microvariables.[8] And indeed, Section 2.3's discussion of the gravothermal catastrophe used statistical mechanical notions of entropy. Furthermore, the collisionless Boltzmann equation and the Fokker-Planck equation for SGS originate from *non-equilibrium SM*...while it is equilibrium SM to which TD putatively reduces. So the inapplicability of TD should be anticipated.

In the next section, I develop the sketch above, of what I take to be the key features of thermodynamics, and then argue that the thermal physics used in SGS cannot be thermodynamics.

4.4 Thermodynamics "Construed"

I will first present my perspective on thermodynamics in general (§4.1), and then argue that thermodynamics so construed, does not apply to SGS (§4.2).

4.4.1 Thermodynamics in general

In this section I rehearse some of the key points from Chapter 2 about how I construe thermodynamics, and add further details about thermal stability.

Recall that equilibrium was a central concept. A system is in a state of thermal equilibrium when its macrovariables no longer vary in time.[9] The state-space of TD is the space of these equilibrium states, parameterised by two (or more) macrovariables, such as p and V for a gas. That a system will reach equilibrium is a presupposition of all of thermodynamics, and is encapsulated by the Minus First Law (Brown and Uffink, 2001).

Once a system reach thermal equilibrium, by the very nature of equilibrium, it will just sit there indefinitely — unless it is intervened upon by an external system, or

[8] Interestingly, away from the Newtonian gravitation regime, matters may be different: the no-hair theorems show that a black hole can be characterised by a few bulk variables: its mass, angular momentum and electric charge.

[9] As discussed throughout this thesis, *absolute* equilibrium is a fiction: it is not realised exactly by any system in the world. Insofar as the systems in question are also described by classical mechanics, then - due to the Poincaré recurrence theorem - we know that given enough time any system will return arbitrarily close to its initial state and thus not remain *indefinitely* in a given state. Thus, Feynman is meant to have quipped that 'equilibrium is the state the system gets into after the very fast stuff [e.g. transients] is over, but the very slow stuff [e.g. Poincaré recurrence] has yet to begin'. But the key point is that we get away with treating a system *as if* they were in thermodynamic, i.e. absolute, equilibrium, because the 'very fast stuff' is over and the 'very slow stuff' does not matter for the phenomena we are interested in.

agent. Throughout Chapter 2, there was an important class of interventions that were represented by a curve through the thermodynamic state space: quasi-static processes.

We saw in Chapter 2 that these quasi-static processes were the source of philosophical controversy. Nonetheless I claimed that the curves in equilibrium state space represent the 'common frontier' of a sequence of small interventions to push the system (approximately) along this curve. The system is 'nudged' from equilibrium (and so out of the equilibrium state space, cf. (Norton, 2016, p. 45)) by altering one of the control variables/external parameters — e.g. by raising the temperature by putting the system (at temperature T_1) in contact with a heat bath at $T_1 + \Delta T$. According to the minus first law of TD, the system will reach a new equilibrium state at this new temperature. The process is then iterated with a series of heat baths at different temperatures, and in this sense the system is pushed along the curve.[10]

This picture of curves in equilibrium space involving nudging the system and it returning to (a perhaps new) equilibrium state requires that the system is thermodynamically stable. In particular, it requires that the second derivative of the entropy is negative[11]: $\frac{\partial^2 S}{\partial E^2} < 0$. In terms of other variables, an alternative requirement for stability is that the heat capacity C_v is positive (cf. Landau and Lifshitz (1969, p.47), Thompson (1972, p. 72)).

This is why positive heat capacity is often taken as a principle of thermodynamics. Indeed, as I mentioned in Section 2.3 (2): strange non-thermodynamic behaviour can occur with a system with negative heat capacity. For example, if a heat bath B (with positive heat capacity, C_v^B) has a lower temperature than a system S with negative heat capacity C_v^S, heat flows from S to B raising the temperature of both. If $|C_v^B|<|C_v^S|$, an equilibrium can be reached where both systems are at a higher temperature than initially.[12]

To sum up: the state space of thermodynamics consists of equilibrium states labelled by a small number of macrovariables. In order for a system to undergo any change, the control variables must be altered by an external system (Lavis, 2017). Thus, a curve through this equilibrium state space does not represent any spontaneous process. Instead, a substantive idealisation is in play: and for this to be connected to the behaviour of real systems the system must be thermodynamically stable so that after small changes in the control variables the system returns to another equilibrium state.

[10] Whether this is an approximation in the sense of Norton (2012) is beyond the scope of this chapter.

[11] Here are some details. To check that the system is not unstable with respect to spontaneously becoming inhomogeneous, (Avoras, 2013, §2.10) imagines splitting the system — already in equilibrium — into two uneven halves (on the left $(E + \Delta E, V + \Delta V, N + \Delta N)$ and on the right $(E - \Delta E, V - \Delta V, N - \Delta N)$ and then he asks: will the entropy increase or decrease? Using $\Delta S = S(E + \Delta E, V + \Delta V, N + \Delta N) + S(E - \Delta E, V - \Delta V, N - \Delta N) - S(E, V, N)$, he shows that in order that $\Delta S = 0$, the entropy must be a concave function of (E, V) at fixed N.

[12] If the converse is true, $|C_v^B|>|C_v^S|$, then the heat bath B will not increase its temperature 'quickly enough' as energy flows in. That is, the system's temperature will increase faster than the heat bath's temperature and so they will not reach the same temperature.

4.4.2 Thermodynamics does not apply to SGS

I claim: the theory discussed above is a far cry from the type of thermal physics used in galactic dynamics, in trying to deal with the 'gravitational million body problem'. In this section, I will argue in a two-pronged attack that thermodynamics — as construed above — does not apply to SGS.

As the state space of TD is the space of possible equilibrium states, we must first consider: what would count as a thermodynamic equilibrium state for a SGS? It is unclear. Binney and Tremaine simply deny that they are any (Binney and Tremaine, 1987, p. 269). Callender is more flexible, suggesting that some unusual states do the job. (However, these unusual candidates are *statistical mechanical* states and so I delay discussion of them until Section 4.5).

The gravothermal catastrophe hints at why: it seems that no equilibrium will be reached since, according to the gravitational potential, the core can keep contracting indefinitely becoming infinitely dense.[13] At this point we face two options.

Either we maintain that such singularities are unphysical, as discussed in Section 4.2.3. They hint at the breakdown of the theory, and any attendant approximations we may have used to deduce the singularity. No such infinities will be reached because other effects will kick in. In the case of globular clusters, the formation of binary stars provides an energy source to resist gravitational collapse. The question is then at which point is the system in equilibrium, which macrovariables are stationary and over which timescales. And there appears to be no clear answer to this. Thus, "the claim that self-gravitating systems have no equilibrium, in particular, is the norm rather than the exception" (Callender, 2011, p. 962) and "galaxies are not in thermodynamic equilibrium" (Binney and Tremaine, 1987, p. 571).

The second option is that the collapsed state of infinite density *is* an equilibrium state — after all, nothing will happen. But I submit that even if we dub this state a thermodynamic equilibrium state, it is only *one* state and to do thermodynamics, we need a whole state space of different possible equilibrium states so that, for instance, we can define curves through this space and so discuss adiabatic and isothermal processes (cf. Section 4.4.1). If we have one lone equilibrium state, then we cannot talk of changing an external parameter such as volume in order to 'nudge' the system to a new equilibrium state, if there is only one state in the whole state space!

This brings me to the second prong of my attack. Even if we could construct an equilibrium state space, there is another problem: SGS are unstable. Perturbing ('poking') the systems, even very gently in the manner required for a quasi static process, can lead to runaway instability. This is unsurprising: even without considering the

[13] "Doing nothing, whilst perhaps difficult for human beings, is altogether excluded for a self-gravitating star system" (Heggie and Hut, 2003, p. 45).

concavity of entropy (i.e. the condition $\frac{\partial^2 S}{\partial E^2} < 0$) and other mathematical conditions (cf. footnote 12), we know this is *exactly* what happens in gravitational systems — recall the 'gravitational clumping' in Figure 4.4! Not only will the system be inhomogeneous with respect to the position of the constituents, but the negative heat capacity implies that an initial temperature gradient is exacerbated. If one system loses energy to the other its temperature increases, whilst the system gaining energy has a decreasing temperature, perpetuating the heat flow between them indefinitely, as seen in Section 4.2.3's gravothermal catastrophe. Initial temperature gradients are accentuated by the heat flow rather than dissipated. This is characteristic of SGS; small inhomogeneities (in the distribution of matter as well as temperature) get amplified not dissipated.

This lack of thermal stability means that after the small interventions used in enacting a 'thermodynamic process', the system will not return to a new (and nearby) equilibrium as required in TD, i.e. by the minus first law of TD. Instead, because SGS do not fulfil the conditions discussed in the previous section (such as positive heat capacity) for thermal stability, a small perturbation will lead to a large change in the state of the system.

Finally, as an aside, notice that the perspective of TD as a control theory brings to light the unthermodynamic nature of SGS. The point here is not merely the obvious, albeit amusing, thought that we cannot manipulate a star, let alone a globular cluster or galaxy. (Cf. the quote from Rowlinson et al. (1993) which I earlier took as emblematic of the Pessimists.) Whilst Elson says 'globular clusters provide an ideal laboratory' (Elson et al., 1987, p. 565), there are important differences between SGS and ordinary TD system that influence how we "manipulate" these systems in computer simulations.[14] In particular, different parameters (such as temperature, density, size of the system) cannot vary independently. The volume of the system is determined by the gravitational potential, and thus volume is not independent of the energy of the system. Hut (1997, p. 10) describes a SGS as having only 'a single coupling constant'— the number of stars. Therefore, even if we could induce the SGS to transition from one state to another, there is only one control variable. As Hut (1997) vividly describes, in a star cluster there are no cylinders or pistons. Instead the stars are confined by their collective gravitational field.

To sum up the argument of this section: I have argued for

The Main Verdict: There is no appropriate equilibrium state space for a SGS. Furthermore, the instability of SGS means that the minus first law of thermodynamics does not apply, and consequently there can be no 'thermodynamic processes'.

[14] For an insight into the difficulties — beyond the sheer size of N — with such computational models, see (Heggie, 2003, p. 83).

4.5 Does statistical mechanics apply to SGS?

I say Yes. That is, the success of thermal physics in application to SGS — such as the collisionless Boltzmann equation etc. — should be attributed to statistical mechanics, not to thermodynamics.

A critic might object that only the mathematical machinery of SM succeeds: stellar systems contain vast numbers of stars (and we can't even solve the 3-body problem!) and thus the calculational problems we faced for a mechanical description of a gas of 10^{23} molecules arise again in the context of self-gravitating systems and so similar mathematical techniques are required. (Indeed, with the good comes the bad: similarities in mathematical success are also followed by similarities in mathematical difficulties. For instance, the scope of SM of SGS is limited to collisionless or weakly interacting systems: some SGS have interactions that are too strong for SM to handle — just like in the case of terrestrial SM! Cf. Callender (2010, §5)).

Accordingly, in this section I discuss the extent to which SM applies. I will agree with the above critic: the applicability of SM does have limitations: in particular, the evolution of self-gravitating systems never reaches a SM equilibrium. Nonetheless, the success of SM is not merely mathematical: when a gravitational kinetic equation can be given, the entropy cannot decrease. Thus, it is not merely the mathematical machinery of SM that applies, although the success of SM must be qualified.

4.5.1 An approach to equilibrium?

Describing *quantitatively* the approach to equilibrium is a key part of the enterprise of SM, known as non-equilibrium SM.[15] Can we describe the behaviour of stellar systems as an approach to equilibrium? If so, this would be a fundamentally statistical-mechanical explanation of the phenomena.

But indeed, there is a problem. Thus Binney and Tremaine say that "we can always increase the entropy of a self-gravitating system of point masses at fixed total mass and energy by increasing the system's degree of central concentration" (Binney and Tremaine, 1987, p.268). Consequently, no density function $f(\overrightarrow{r}, \overrightarrow{v}, t)$ maximises entropy for finite mass and energy.[16] So if SM equilibrium is taken to be defined as the maximum entropy state (a feature common to both the Gibbsian and Boltzmannian definitions of equilibrium) a SM equilibrium cannot be found for finite systems.

There are three possible reactions. First, Binney and Tremaine conclude that the

[15] Thermodynamics only states that a system will go to equilibrium (a tendency that has been dubbed the "minus first law" of thermodynamics, as noted in Section 4.1). It does not say anything about how fast equilibration occurs.

[16] A density function that maximises entropy exists only for the isothermal sphere which has infinite mass and energy (Binney and Tremaine, 1987, p. 268).

behaviour of SGS cannot be treated as a relaxation to equilibrium. $f(\overrightarrow{r}, \overrightarrow{v}, t)$ is not analogous to the velocity distribution of an ideal gas relaxing to the Maxwell-Boltzmann distribution. Galaxies and other typical stellar configurations are not the result of a long-term thermal equilibrium (Binney and Tremaine, 1987, p. 269).

Second, Callender (2011) takes a different view and suggests that the unconventional states such as the collapsing core-halo states and similar Dirac δ-function 'singular peaks' should "be regarded as equilibrium states for the same reasons cups of coffee at room temperature can be" (Callender, 2011, p. 968). If these states can be interpreted as SM equilibrium states, then (on this view) the behaviour of SGS could be an approach to equilibrium.

However, a cup of coffee is in a *local* equilibrium state. Rather than being described by a global Maxwellian distribution such as,

$$f(p,q) = Ne^{\frac{p^2}{2mkT}}, \tag{4.7}$$

the system is in a local Maxwellian distribution. For instance, the temperature of the coffee varies with position, but locally looks like an equilibrium state. Thus over certain distance scales, the coffee looks like it is in equilibrium.

$$f(p,q) = N(q)e^{\frac{p^2}{2mkT(q)}} \tag{4.8}$$

Whilst the cores of stars are in local but not global equilibrium, the unconventional states that Callender proposes are not states of local equilibrium.[17] Further, by Callender's own lights the 'Dirac peak' is the wrong state to use; since he claims that the canonical ensemble (in which the state is defined) is the wrong ensemble to use for astrophysical systems and instead the microcanonical ensemble should be used (Callender, 2011, p. 967).

Thirdly, you can walk straight in a particular direction without reaching some prescribed destination. That is: when a gravitational kinetic equation is given, the entropy (as a function of the distribution function) increases but never reaches a maximum.

[17] Individual stars are an interesting case: are they examples of SGS that admit of a TD description (and so provide a counterexample to my thesis)? Since they are in local equilibrium, perhaps some of the problems I raised in Section 4.4.2 for SGS such as globular clusters (i.e. 'no equilibrium states') do not apply. Yet the equilibrium state in equation 4.8 is a probability distribution over microvariables p and q — which, according to my classification of Section 4.3, is an example of SM, rather than TD, machinery. What to conclude? I think it will depend on how one draws the line, i.e. what physics one designates as 'thermodynamics'. Thus, at the outset of his classic monograph 'Thermodynamics of the Stars', E.A. Milne distinguishes two senses of thermodynamics: (i) a 'restricted sense as denoting the study of the equilibrium states of enclosed systems' and (ii) a 'general way to denote the study of all those phenomena in which temperature plays a part... the science of heat transfer' (Milne, 1930, p. 5). A more permissive definition of TD — along the lines of Milne's (ii) — may classify this case as TD. But for my account propounded in this chapter, individual stars seem to be a boundary case — and one deserving of further study.

Were there a maximum entropy state, we might want to call this the SM equilibrium state. For SGS there is no such destination, but nonetheless these systems head in that direction: so I conclude that there is (a weak sense) in which they *approach* equilibrium, although they do not reach it.

This meshes nicely with our earlier discussion in Section 4.4.2. There we saw (in the 'second prong of attack') that due to the instability of SGS, the minus first law of TD does not hold: SGS do not spontaneously return to TD equilibrium after small perturbations. I cannot of course go to into detail here about the exact relations between SM equilibrium and TD equilibrium. But we expect non-equilibrium SM to in some way justify or underpin the minus first law of TD. So not reaching SM equilibrium (and consequently limiting the applicability of SM) fits with my earlier conclusion that the minus first law does not apply to SGS. Furthermore, as we saw in the 'first prong of attack', SGS do not reach a state of thermodynamic equilibrium. Had SM equilibrium states been available this may have suggested that there is a way to find a TD equilibrium state space, since SM equilibrium is meant to (in some sense) ground TD equilibrium. Thus, the lack of SM equilibrium further supports the conclusion of Section 4.4.

4.5.2 Surprise?

Should we think it surprising that statistical mechanics applies here? I say: No. The application of SM to SGS is not surprising. For very similar assumptions are used in the descriptions of dilute gases and of the SGS that SM is capable of describing. First, the collisionless Boltzmann equation assumes that the stars in the system are intrinsically identical (in particular having the same mass m) and non-interacting—just as Boltzmann assumed for an ideal gas. Secondly, an assumption similar to the Boltzmann's infamous 'Stosszahlansatz' is made: the presence of star 1 being found in a particular area of phase space does not raise or lower the probability of star 2 being found there (Binney and Tremaine, 1987).

The application of SM may not be surprising, but it might nonetheless still be surprising that we seem to have an SM description *without* a TD description. After all, the concepts of each theory are frequently assumed to be intertwined: as discussed in section 4.3, the two theories are not always cleanly separated. But in the next section I given an explanation of why no TD behaviour emerges from the SM description.

4.6 The Bottom-Up Explanation

I have concluded that whilst there is (to some extent) a SM description of SGS, thermodynamics does not apply. This conclusion allows a peace to be made between the

Pessimists and the Optimists. The Pessimists were sceptical that a theory concerned with steam engines could be extrapolated so far from its human origins to such exotic realms. To the extent that they are talking about the applicability of *thermodynamics*, they are correct. But the Optimists are correct too — thermal physics *is* successful in these exotic gravitational realms — provided that by thermal physics we mean *statistical mechanics*.

But this verdict might seem surprising, given that in Chapter 2, I claimed that TD is reduced to SM: that is, the lower-level theory T_b, SM, captures the behaviour of systems described by T_t, TD. How can the two theories come apart? Throughout this chapter, I've argued that SGS do not exhibit the equilibrating behaviour that is encoded in the Minus First law. This behaviour is a prerequisite for a system to represented by an equilibrium state-space as in TD . But SM has slightly different concerns. In particular, a large part of the SM enterprise is to quantitatively describe the approach to equilibrium; indeed, all of non-equilibrium SM is about this. Thus, there is a sense in which the two theories have different concerns, i.e. different subject matters.

In this section, I sketch an explanation of this 'coming apart' of SM and TD in the case of SGS, which I call the bottom-up explanation. The explanation of this fact is that a particular mathematical limit — the thermodynamic limit — does not exist (Padmanabhan, 1990, p. 295). The topic of the thermodynamic limit is a vast one; see Ruelle (1999) for a classic presentation of both continuous and discrete systems. The key idea is whether there is a mathematically well-defined ideal infinite system obtained by $n \to \infty$, where n is the number of constituents of a system.[18] Usually, the thermodynamic limit not only takes $n \to \infty$ but also fixes the density, i.e. $\frac{n}{V} \to k$, whilst $n \to \infty$. Generically, in the thermodynamic limit, a TD description is recovered from the SM description. Thus, the bottom-up explanation I advocate is:

The Bottom-Up Explanation: Generically, in the thermodynamic limit, a TD description is recovered from the SM description. But the thermodynamic limit does not exist for self-gravitating systems.

In filling out this explanation, I will discuss: (1) why the limit does not exist and (2) the significance of its not existing.

(1) The thermodynamic limit does not exist for self-gravitating systems. To prove the existence of a thermodynamic limit, it suffices to show that two conditions are met by the system under consideration (Penrose, 1979, p. 1963); see also (Thompson, 1972, ch. 3) and (Ruelle, 1999, ch. 3):

[18] I should note that how such limiting procedures should be understood and classified is the topic of philosophical debate, e.g. Norton (2012).

1. *Tempering*: the interaction between distant constituents must be negligible. Here is a simple example of such a tempering condition: a pair potential $U(x)$ (where x is the distance between the two particles) has a finite range if there is a distance R_0 such that $U(x) = 0$ for $|x| \geqslant R_0$.

2. *Stability*: the interaction is stable: there is a real number $B \geqslant 0$ such that $\forall n$ the potential energy of n constituents located at any spatial points $x_1, ...x_n$, $U(x_1, x_2...x_n) \geqslant -nB$.

These two conditions are violated by the long-range and short-range nature, respectively, of the gravitational potential.

1. *Tempering* is violated because the gravitational potential between two distant particles (i.e. stars) is $V \sim \frac{1}{r}$, which is unlike the potential above: whilst the potential decreases with distance, $U(x) \neq 0$ for any $|x|$. As we saw earlier, the interaction between a star and its distant neighbours is not negligible, but indeed quite the reverse: as seen in Figure 4.1, the long-range nature of the gravitational potential dominates the behaviour of the cluster. Of course what counts as 'negligible' is not categorical, but for no degree of accuracy does it seem we can treat these gravitational interactions as 'negligible'.

2. *Stability* has two components. Firstly, the potential energy must be bounded from below; condition 2 states that for n constituents there must be a number, $-nB$, such that the potential energy is always greater than this number. But, as seen in Figure 4.2 the gravitational potential is not bounded from below: as $r \to 0$, $U \to -\infty$. Secondly, *Stability* requires that U does not grow faster, as a function of n, than n. This too is violated by the gravitational potential. Hut (1997, p. 8) calls the gravitational potential 'superextensive': in fact, $U \sim n^{\frac{5}{3}}$.[19]

Thus, SGS differ from ordinary thermodynamic systems in (at least) two respects: the thermodynamic limit does not exist and energy is not extensive for SGS.[20]

[19] The 'back of the envelope' justification Hut gives for this is as follows. "Take a large box containing a homogeneous swarm of stars. Now enlarge the box, keeping the density and temperature of the star distribution constant. The total mass M of the stars will then scale with the size R of the box as $M \propto R^3$, and the total kinetic energy E_{kin} will simply scale with the mass: $E_{kin} \propto M$. The total potential energy E_{pot}, however, will grow faster: $E_{pot} \propto \frac{M^2}{R} \propto M^{\frac{5}{3}}$. Unlike intensive thermodynamic variables that stay constant when we enlarge the system, and unlike extensive variables that grow linearly with the mass of the system, E_{pot} is a superextensive variable, growing faster than linear" Hut (1997, p. 8).

[20] Interestingly, despite their surface similarities, the gravitational potential and Coulomb potential differ: the thermodynamic limit has been proven to exist for certain electromagnetic systems (cf. Lieb and Lebowitz (1972), (Thompson, 1972, p. 71)). But this proof requires quantum considerations (roughly, that there is a ground state so that the energy is bounded from below, so that the stability condition is satisfied. The tempering condition can also be satisfied by electromagnetic systems. The presence of both positive and negative charges leads to Deybe shielding (Callender, 2011, p. 961): the force due to distant charges is 'shielded' so that $U(x) \approx 0$ for suitably large $|x|$. But there is no analogous effect of Deybe shielding for gravitational systems because the gravitational force does not 'saturate' (Lévy-Leblond, 1969).

(2) What is the significance of the lack of limit?

In full generality, this is a hard question to answer. But here our task is smaller: to explain why an SM description of SGS is successful, but yet no TD behaviour emerges. Given how the two theories are seemingly interwoven, how do we have the applicability of one without the other? The answer is that the thermodynamic limit connects the two theories, but because the limit does not exist for SGS, we should be less surprised at the applicability of one without the other.

I shall briefly spell out three examples of the thermodynamic limit connecting SM and TD. Here the idea is that the differences between the TD description and the SM description are washed out in this limit:[21]

(A) The thermodynamic limit is usually used to reveal features of SM functions. For instance, the canonical and microcanonical ensembles are equivalent in the thermodynamic limit. The significance of this result is sometimes glossed as: the equivalence of ensembles in the limit shows that the same thermodynamic functions (and so behaviour) results — no matter which ensemble is used to calculate those functions.

(B) SM descriptions involve probabilities, whereas the TD descriptions are non-probabilistic. But in the thermodynamic limit, the probability of fluctuations away from the mean value (e.g. the mean energy), tend to zero. Thus, in the limit, the SM description becomes more akin to the non-probabilistic TD description.

(C) Furthermore, only in the thermodynamic limit are certain SM quantities extensive. For instance, the Gibbs free energy is only extensive in this limit. Or, as seen in section 4.2.3, $H_{12} = H_1 + H_2$ is only an approximation, even for short-range potentials present in familiar cases. Yet, the energy of a subsystem is proportional to the volume, whereas interactions between subsystems are proportional to the boundary's surface area — and so scaling means that interactions thus become negligible, provided the subsystems are big enough. Thus strictly speaking, even for short-range potentials, entropy and energy are only extensive in the thermodynamic limit.

Because some SM quantities are only truly extensive in the thermodynamic limit, they are only identical to their TD correlates in this limit. Thus, it is usual to say that in the thermodynamic limit, the thermodynamic formalism/functions are recovered from the SM description. For example, Oliver Penrose claims that "the first objective of the study of the thermodynamic limit is to demonstrate that in this limit, the laws of thermodynamics apply and to justify our statistical mechanical recipes for calculating thermodynamic functions..." (Penrose, 1979, p.1957).

But is the existence of the thermodynamic limit a necessary and/or a sufficient condition for recovering a TD description from a SM description? Callender (2011, p. 975) suggests it is neither necessary nor sufficient.

[21] But of course, there is much more work to be done to understand how the limit connects SM and TD; here I only give a suggestive sketch.

Yet it seems like the existence of the TD limit suffices for a TD description to be recovered from the SM description. If it is a sufficient condition for the applicability of TD, this might help determine which SM systems also fall under the purview of TD. But it would be less enlightening for ruling systems *out* — since there could exist an alternative sufficient condition for the applicability of TD. Furthermore, analysing the counterfactual claim that 'had the TD limit existed, then TD would have applied' is hard for SGS: for the TD limit to exist, the system would have been fundamentally different — as I argued in Section 4.2, the form of the gravitational potential dominates the behaviour of SGS. And it is this potential which fails the tempering and stability conditions.

But I agree with Callender that the TD limit is not established to be a *necessary* condition: there *could* exist a type of system for which the thermodynamic limit of a SM description does not exist but to which TD nonetheless applies.[22] Thus, the thermodynamic limit is not a 'blanket prescription': one need not prove the existence of the thermodynamic limit, in order to be licensed to use the laws of thermodynamics. Instead, we apply the ideas of thermodynamics to a system just when it is useful to do so. Thus, thermodynamics might be useful for a system for which the limit does not exist. (Black holes are a putative example — but much controversy surrounds this claim.) But I contend, contra Callender, that SGS do not give us this counterexample, since — as I hope to have established in Section 4.4.2 — TD is inapplicable to SGS.[23]

To sum up: the fact that the thermodynamic limit does not exist for SGS explains why the applicability of SM and TD come apart for these systems.

Whilst I have been careful throughout the preceding section to talk of TD being 'recovered' from SM in the limit, the question I must now face is: how does the main focus of this chapter— the domain of applicability of thermodynamics — relate to the reduction debate?

[22] Of course, it would be an interesting development if the existence of the thermodynamic limit was a necessary condition for the applicability of thermodynamics — but this is not something I can establish here.

[23] Whilst it is conceivable that there is a system for which the thermodynamic limit does not exist and yet thermodynamics is useful, there are reasons to be confident of the importance of 'large N' in what is after all, a science of the *bulk* properties of matter. For instance, according to the Boltzmannian perspective on SM, systems inevitably approach equilibrium because of the overwhelming vastness of the equilibrium macrostate compared to the other macrostates. But this requires N to be large — otherwise there will not be one macrostate dominating the available phase space. Whilst not required for the cogency of their account in quite the same way, the Gibbsian also depends on the thermodynamic limit/large N. For, as mentioned above, Gibbsian ensembles are only equivalent in the thermodynamic limit – a fact held as vital by physicists. For instance, Huang says "From a physical point of view, a microcanonical ensemble must be equivalent to a canonical ensemble, otherwise we would seriously doubt the utility of either" (Huang, 1987, p.148): cited in Callender (2011, p.977). Thus, whilst the Gibbsian canonical ensemble is applicable to a one-molecule gas — unlike the Boltzmannian picture — the thermodynamic limit is nonetheless seen as crucial (Thompson, 1972).

4.6.1 The connection to reduction

The first connection to reduction debate is as follows: there is a debate about whether the role of the thermodynamic limit in SM descriptions of phase transitions blocks the reduction of TD to SM. Briefly, the concern is that a singularity in the free energy can only be achieved in the infinite limit[24] and this infinite idealisation is unrepresentative of real systems and so some contend that a complete reduction to SM is unavailable. Others argue that the use of the limit need not block reduction and the relevant 'singular type' behaviour is seen 'on the way' to the limit (Butterfield, 2011b). Does the above discussion reveal that the thermodynamics limit is crucial (in a way problematic for reduction) even away from the contentious case of phase transitions? Should it be worrisome that functions such has the Gibbs free energy are only extensive (and thus like their TD counterparts) in the thermodynamic limit? I think not. Rather than a qualitative difference that springs out only at the limit — causing water to boil and other phase transitions to occur — nothing so glamorous happens. Rather, here the situation is one of 'mathematical tidiness', i.e. the interaction Hamiltonian *really* is $= 0$, rather than ≈ 0 . Indeed, I contend that if reduction were thwarted by the 'exactness' only existing in the 'unrepresentative' limit, then this would set the bar so high for inter-theoretic reduction that few or no cases would pass it.

Indeed, this was the intuition behind my functionalist account in Chapter 2. The three examples (A)-(C) above showed how a TD description was recovered from the SM description in the limit: we get back the extensivity of energy and categorical, rather than probabilistic, functions. But I submit that the TD limit is not essential for reduction (contra Batterman (2001)). Whilst the TD limit washes certain differences between the SM and TD descriptions of a given system, often these differences do not matter for reduction. This is because our SM descriptions successfully capture the TD behaviour *before* the limit. (This was the project of Chapter 2). In fact, if thermodynamic behaviour only occurred in this limit, we would never expect to see TD behaviour because no actual systems are infinite. Thus, I take it to be a positive feature of my account that the TD limit is not essential.

I now connect the question of domain of applicability of thermodynamics to the reduction debate. Had TD applied to SGS, this could have been used to support the view that 'thermodynamics is fundamental', in the manner of Eddington's and Planck's views (cf. Section 4.3).[25] But instead, this case study arguably adds to the conceptual priority of SM. That SM applies without the emergence of TD behaviour agrees with the moral of Wallace (2015a); SM is foundationally important *not only* insofar as it is

[24] The free energy is $F = -kTlnZ$ where Z is the canonical partition function $Z = \Sigma_n exp(-\frac{E_n}{kT})$. In order for there to be a non analyticity in this function, $n \to \infty$.

[25] This veneration of thermodynamics, as seen in the quote by Einstein in the introduction to this thesis, is one reason to think that the relationship between TD and SM is not just one of horizontal reduction.

connected to TD.

Of course, the more charitable interpretation of the 'TD is fundamental' view is that it is merely stressing the importance of TD: in particular, one might read the Eddington quotation in Section 4.3 as emphasising the epistemic security of TD. The claim is that the principle of entropy increase is a principle for which we have vast amounts of evidence; in part because the domain of applicability of thermodynamics is taken to include everyday occurrences such as people ageing, buildings crumblings and coffee cooling. (Such a universal scope is also part of Albert and Loewer's 'Mentaculus' project, cf. Albert (2000); Loewer (2018)). But the case of SGS heeds us to be cautious: the scope of TD is not universal.[26]

4.7 Conclusion

The detailed empirical success of our descriptions of SGS form a fascinating, often stunning, part of physics. But it is a success to be credited to the framework of ideas provided by statistical mechanics, not thermodynamics. Thus, the situation is: there is a SM description of SGS such as globular clusters and elliptical galaxies, but no thermodynamic behaviour emerges. The unusual unstable behaviour of SGS, negative heat capacities and runaway instabilities, is alien to thermodynamics — but this is unsurprising when we consider the principle theory of thermodynamics as a control theory whose state space is that of equilibrium states. In contrast, the constructive theory of SM applies to SGS; we can write down a probability distribution for a given star to occupy a certain position and have a certain velocity and that entropy associated to that distribution is non-decreasing. The applicability of SM without TD has a bottom up explanation: the thermodynamic limit does not exist for SGS and there is no mathematical bridge between these two theories.

[26] This deflationary position about the grand claims surrounding TD meshes with my discussion of the Second law in Chapter 2.

Conclusion

In this concluding Chapter, I will suggest (1) how my conclusions throughout this thesis tie together, (2) what the potential implications for other case studies in philosophy of physics are, and (3) in which ways the issues considered here are interwoven with the wider philosophical landscape.

(1) In Chapter 1, I distinguished vertical reduction from horizontal reduction. For vertical reduction, T_t abstracts away from the details described by T_b: as such, the two theories often describe different subject matters. In Chapter 3, this was clearly demonstrated by the vertical reduction of statistical mechanics to its underlying dynamics. The underlying quantum or classical mechanics describe the microscopic details, whereas statistical mechanics concerns the macroscopic patterns that result from those details. Indeed, that the higher-level irreversible equations describe something different from the microscopic equations was part of the justification of coarse-graining: coarse-graining abstracts away from the microscopic details. We should not reject coarse-graining, because then we would lose our descriptions of these higher-level patterns: and thereby useful information about transport coefficients and the equilibration timescale. These equations' descriptions of the approach to equilibrium is the statistical mechanical underpinning of the Minus First Law of TD.

For horizontal reduction, T_t approximates T_b — perhaps to a certain degree of accuracy, or for a particular domain of applicability. An older theory T_t, where it was successful, approximates the better, successor theory T_b. This is demonstrated by the relationship between Newtonian mechanics and special relativity: in the domain where the velocities are much smaller than the speed of light (such that $\frac{v}{c} \rightarrow 0$), Newtonian mechanics approximates special relativity.

But this distinction between vertical and horizontal reduction was one of degree, not of kind, since approximation and abstraction can overlap. One of the conclusions of Chapter 2 is that the relationship between thermodynamics and statistical mechanics is not just a horizontal reduction. Instead, there are elements of vertical reduction, since the two theories have (to some extent) different subject matters. The different subject matters were manifest when considering the dynamics of each theory. For example, quasi-static processes which I argued were *so* central to thermodynamics (especially the second law), are not so central in SM.

What are the philosophical consequences of the two theories having different subject matters? It suggests that SM is not simply the successor to TD: this is not a straightforward horizontal reduction. But this conclusion is perhaps to be anticipated given the high esteem in which some physicists hold thermodynamics. We saw in Chapter 4 that Eddington venerated thermodynamics, and in the introduction, we saw that Einstein claimed that the laws of thermodynamics are applicable to all systems and would never be overthrown. Such claims would seem odd or surprising if thermodynamics were merely a superseded theory.

But one might wonder how these grand claims about thermodynamics fit with my construal, throughout this thesis, of thermodynamics. In particular, Einstein, Planck and other giants claim that thermodynamics has a certain universality: how does this fit with my verdict in Chapter 4 that thermodynamics does not apply to SGS? Or my claim in Chapter 2 that the second law is not the source of all time-asymmetry?

My construal of thermodynamics was narrow: and so, one might read those who venerate thermodynamics as referring a wider set of ideas than my construal of 'thermodynamics'. For example, Milne (1930) distinguishes (i) the narrow sense I have endorsed from (ii) the wider meaning of 'thermodynamics' as the science of all phenomena involving temperature. So in the passages venerating thermodynamics, the term 'thermodynamics' may have Milne's second sense: 'thermodynamics' may refer to all of thermal physics.

There is, of course, a connection between Einstein's claim about (i) inviolability of the laws of thermodynamics and (ii) the universal scope of these laws, as follows. When one discovers a putative violation of a given law, one might restrict the scope of that law, so as not to be applicable to that type of system or situation. For example, if contra Newton's original intention, we claim that the laws of Newtonian mechanics are only applicable to systems with suitably low velocities, then we might be able to avoid violations of these laws.

And so, for the case of the thermodynamics, one might likewise restrict the scope of the TDSL, in light of certain violations. But, in Chapter 2, I argued that this is *not* what is going on in the case of thermodynamics. I have not limited the scope of thermodynamics in light of its reduction to statistical mechanics. Some say that the TDSL is only a statistical truth: that there are violations. For example, we saw that Maxwell restricted the scope of the TDSL to large systems — in light of the molecular nature of matter. But I argued that the nature of matter does not have the consequence that we can *reliably* have an engine more efficient than a Carnot engine. Thus, my project in part supports Einstein's and others' claims that the laws of TD are not violated.

However, I do not endorse the assumption of universality. It is important to notice how central quasi-static processes are to thermodynamics. And this, along with the importance of an equilibrium state space, motivated my 'narrow scope' reading of thermodynamics. Thus, my narrow scope reading of thermodynamics was not due to issues of reduction — but instead, due to the internal details of thermodynamics. Nevertheless, I think my stricter construal of the term 'thermodynamics' is fruitful: it sheds light on the debate about self-gravitating systems. I ruled that thermodynamics was inapplicable to SGS (although SM was applicable to a certain extent). Thus, I claim that thermodynamics does not have an unrestricted scope. But that is not to say that its domain cannot be extended from steam engines.

One advantage of my functionalist approach to thermodynamics, is that it puts the focus squarely on the behaviour of the system. For example, whilst I ruled that SGS are not in the domain of applicability of thermodynamics, I am open to the idea that black holes may be truly thermodynamic objects. This leads to (2).

(2) There is an analogy between the laws of black holes, and the laws of thermodynamics. Loosely, the area of a black hole is non-decreasing and this is akin to the non-decreasing thermodynamic entropy.[27] Physicists, such as Wald, claim that "the laws of black hole mechanics *are* the laws of thermodynamics" (Wald, 2001). But this identity claim has generated controversy in the philosophy of physics. Some, such as Dougherty and Callender (2016) and Maudlin (2017) are sceptical. There are various differences between thermodynamic quantities, such as temperature, and their black hole correlates, for temperature: surface gravity. For example, temperature is intensive but surface gravity is not. Yet others are optimistic (cf. Wallace (2017, 2018), Prunkl and Timpson (2018), Curiel (2014)). For example, Prunkl and Timpson (2018) show that a black hole can be part of a Carnot cycle, and thus interact with other systems in a suitably thermodynamic manner. Consequently, they conclude that the Bekenstein entropy is the thermodynamic entropy: $S_{TD} = S_{BH}$.

But I think this identity claim should be read as a realisation claim: the thermodynamic entropy S_{TD} is realised by the Bekenstein entropy S_{BH} (which is related to the black hole's area), and likewise, the thermodynamic temperature is realised by the black hole's surface gravity. If surface gravity only has to play the role of temperature, then the fact that there are differences need not matter. The scepticism of Dougherty et al. can be defused in the same way that, in Chapter 2, I defused Sklar's worries about statistical and non-statistical quantities. Namely: these differences are not part of the functional or nomological role of temperature. Whether this functionalist strategy is successful in the case of black holes depends on whether they display TD behaviour: a claim which I postpone for future work.

(3) Having considered how my conclusions relate to other issues in philosophy of physics, I now consider the connections to the wider philosophical landscape. In Chapter 3, *abstraction* was centre stage in justifying coarse-graining in SM. But abstraction was also central in Chapter 1's overarching framework of levels of description. For

[27] I say this is loose because, as we saw in Chapter 2, the TD entropy can decrease if the system is not isolated.

both Chapter 1's general idea of abstraction, as well as Chapter 3's specific cases in SM: there are many different possible abstractions. Abstracting involves throwing away lower-level details, and so there is a lot of choice or variety over which details get thrown away. In List's framework, there are many possible abstraction maps σ, i.e. partitions, over the lower-level possible worlds. In other words, there are so many possible higher-level variables. As I noted in Chapter 1, somewhere in Platonic heaven there is a mathematical function describing the trajectory of the centre of mass of Trump's hand and my cat's claw. And, as we saw in Chapter 3, there is a higher-level variable, the centre of mass of all philosophers of physics. Whilst, in principle, there is an uncountable plurality of such variables, we do not use them. We would be hard pressed to write down the function for the trajectory of such centres of mass. In general, we describe the world using variables that we can manipulate, measure and control. In the specific cases in SM, we want an abstraction map which leads to an autonomous level of description: the lower-level details really do not matter for the evolution of the higher-level variables.

Abstraction was important — both for giving a detailed justification of coarse-graining in SM — but also more generally in thinking about the relationships between different levels of description. Throwing away details, i.e. not caring about certain things, in part explains why the higher-level theory has a different subject matter. Is this an anthropocentrism? It does tie the different levels of description to our cognitive abilities and our epistemic standpoint on the world. But, as I discussed in Chapter 3 for SM, and in Chapter 2 for TD, this is a very *general* anthropocentrism — it is not one specific to thermal physics. As such, it raises questions about the plausibility of metaphysical realism. But whether our descriptions of reality are really mind-independent is, unsurprisingly, beyond the scope of this thesis. But I do claim to have established that thermal physics is not anthropocentric in a way different from other scientific theories. So we don't need to be *selective* in our scientific realism.

In Chapter 2, I discussed how processes in thermodynamics require interventions by an external agent. But interventions need not be considered anthropocentric — it doesn't matter that *we* stir or shake the system, only that some external system does. Indeed, away from fundamental physics, sciences such as chemistry are awash with interventions.

In Chapter 3, I argued that the choice of a particular coarse-graining operator $\hat{P}$ doesn't depend on our whims, but rather on whether we can find autonomous higher-level dynamics (which I called the C^+ dynamics). These higher-level dynamics were autonomous provided that they are 'forwards-compatible': the dynamical evolution of the coarse-grained probability distribution ρ_r is independent of the lower-level details (in this case: ρ_{ir}).

I hope this point about autonomy will generalise. If the higher-level description, theory or equation is independent of, or autonomous from, certain (but clearly not all) lower-level details, then it will be robust under changes of these irrelevant details. The lower-level states (or possible worlds) that differ only in these irrelevant details will form an equivalence class — the members of this class will multiply realise the higher-level state.

This explains the sense in which the higher-level patterns described by the special, i.e. non-fundamental, scientific theories are robust. If the higher-level variables were not autonomous from the lower-level details — e.g. if psychology were sensitively dependent on the advances and details of cosmology — then we would struggle to have the successful special scientific theories that we do. Indeed, that we are lucky to hit upon different robust patterns at different scales or levels is the reason, to answer Loewer (2009), why there is anything except physics.

I have described a sense in which the higher-level theories might be independent, or autonomous; for example, if we have the meshing dynamics situation of Chapter 3. But I also submit that these higher-level theories often describe novel phenomena. The subject matter of psychology is different from the subject matter of physics. The higher levels care about different things, and this is why a completed physics will not tell us anything about minds and society (pace Putnam and Oppenheim). Even if black hole thermodynamics guides us to a theory of quantum gravity, and even if this theory were a final theory of 'completed physics'— we will still be in the dark about psychological phenomena like stereotype threat, economic phenomena like the 2008 financial crash or why male platypuses have venomous spurs. But we needn't look at disparate levels of description: the subject matter of statistical mechanics, i.e. the concerns of the theory and what it aims to predict, are different from the underlying microdynamics.

Thus, I claimed that if the higher-level phenomenon that T_t describes, is both robust and novel, then —in this, admittedly weak, sense— it is emergent. And this is compatible with T_t being reduced to T_b. I hope that this general claim from Chapter 1 has been convincingly demonstrated in Chapter 3.

As a general methodological lesson, I hope to have demonstrated the dangers of not engaging with the pernickety details in physics for metaphysics. In considering the case studies of reduction in thermal physics, I have engaged with the details of a range of debates in thermal physics, and gone beyond considering the ideal gas. I submit that there is a danger for philosophy beyond the niche of philosophy of physics, in considering the ideal gas to the exclusion of other cases: it makes the physics look too simple.

Darwin said "false facts are highly injurious to the progress of science, for they often endure long..." (The Descent of Man as cited by Wilson (1985)). To echo Wilson: the false fact of $T = \langle K \rangle$ — and I would add: the overemphasis on classical SM to the exclusion of quantum SM — has endured too long.

And this has been injurious to the progress of *philosophy*, for it paints the physics in too monochrome a light. By appreciating the richness of the physics, we see that many of the features — multiple realisability, autonomy and emergence — are present in the relationships between physical theories; those very theories that are meant to provide the unproblematic contrast class to the philosophically interesting, or problematic, cases of pain, the mind or the special sciences.

Bibliography

A. Greven, G. Keller, G. W. (2014). *Entropy*. Princeton University Press.

Albert, D. Z. (2000). *Time and chance*. Harvard University Press, Cambridge, Mass.

Alexander, S. (1920). *Space, time and deity: the Gifford Lectures at Glasgow, 1916-1918; in 2 volumes*. Macmillan, London.

Atkins, P. (2007). *Four laws that drive the universe*. Oxford University Press.

Avoras, D. (2013). Lecture Notes on Thermodynamics and statistical mechanics (a work in progress). San Diego lecture notes, available at http://courses.physics.ucsd.edu/2010/Spring/physics210a/LECTURES/210COURSE.pdf.

Baierlein, R. (1971). *Atoms and information theory: An introduction to statistical mechanics.* WH Freeman San Francisco.

Balescu, R. (1997). *Statistical Dynamics: Matter out of Equilibrium.* Imperial College Press, London.

Balescu, R. (2005). *Aspects of anomalous transport in plasmas*. CRC Press.

Batterman, R. (2001). *The devil in the details: Asymptotic reasoning in explanation, reduction, and emergence*. Oxford University Press.

Batterman, R. (2010). Reduction and renormalization. *Time, Chance, and Reduction: Philosophical Aspects of Statistical Mechanics*, pages 159–179.

Batterman, R. W. (1995). Theories between theories: Asymptotic limiting intertheoretic relations. *Synthese*, 103(2):171–201.

Batterman, R. W. (2009). On the explanatory role of mathematics in empirical science. *The British Journal for the Philosophy of Science*, 61(1):1–15.

Bedau, M. A. and Humphreys, P. E. (2008). *Emergence: Contemporary readings in philosophy and science.* MIT press, Cambridge, Mass.

Bekenstein, J. D. (1973). Black holes and entropy. *Physical Review D*, 7(8):2333.

Bergmann, P. G. and Lebowitz, J. L. (1955). New approach to nonequilibrium processes. *Physical Review*, 99(2):578.

Berry, M. V. (2002). Singular limits. *Phys. Today*, 55.

Binney, J. and Tremaine, S. (1987). *Galactic dynamics*. Princeton University Press, Princeton, New Jersey.

Blatt, J. (1959). An alternative approach to the ergodic problem. *Progress of theoretical physics*, 22(6):745–756.

Block, N. (2003). Do causal powers drain away? *Philosophy and Phenomenological Research*, 67(1):133–150.

Blundell, S. J. and Blundell, K. M. (2009). *Concepts in thermal physics*. OUP Oxford.

Brandao, F. G., Horodecki, M., Oppenheim, J., Renes, J. M., and Spekkens, R. W. (2013). Resource theory of quantum states out of thermal equilibrium. *Physical review letters*, 111(25):250404.

Bridgman, P. W. (1943). *The nature of thermodynamics*. Harvard University Press, Cambridge, Mass.

Broad, C. D. (1925). *The mind and its place in nature*. Routledge, London.

Brown, H. R. (2005). *Physical relativity: Space-time structure from a dynamical perspective*. Oxford University Press, Oxford.

Brown, H. R. and Uffink, J. (2001). The origins of time-asymmetry in thermodynamics: The minus first law. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 32(4):525–538.

Brush, S. G. (1976). *The Kind of Motion We Call Heat. A History of the Kinetic Theory of Gases in the 19th Century. Book 1: Physics and the Atomists. Book 2: Statistical Physics and Irreversibel Processes*. Amsterdam.

Butterfield, J. (2011a). Emergence, reduction and supervenience: a varied landscape. *Foundations of Physics*, 41(6):920–959.

Butterfield, J. (2011b). Less is different: emergence and reduction reconciled. *Foundations of Physics*, 41(6):1065–1135.

Butterfield, J. (2012). Laws, causation and dynamics at different levels. *Interface Focus*, 2(1):101–114.

Butterfield, J. and Bouatta, N. (2012). Emergence and reduction combined in phase transitions. In *AIP Conference Proceedings 11*, volume 1446, pages 383–403. AIP.

Callender, C. (1999). Reducing thermodynamics to statistical mechanics: the case of entropy. *The Journal of Philosophy*, 97(7):348–373.

Callender, C. (2001). Taking thermodynamics too seriously. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 32(4):539–553.

Callender, C. (2004). There is no puzzle about the low entropy past. In Hitchcock, C., editor, *Contemporary Debates in Philosophy of Science*, chapter 12. Wiley-Blackwell, Oxford.

Callender, C. (2010). The past hypothesis meets gravity. In Gerhard Ernst, A. H., editor, *'Time, chance and reduction: philosophical aspects of statistical mechanics'*, pages 34–58. Cambridge University Press, Cambridge.

Callender, C. (2011). Hot and heavy matters in the foundations of statistical mechanics. *Foundations of Physics*, 41(6):960–981.

Carathéodory, C. (1909). Untersuchungen über die grundlagen der thermodynamik. *Mathematische Annalen*, 67(3):355–386.

Cartwright, N. (1999). *The dappled world: A study of the boundaries of science*. Cambridge University Press, Cambridge.

Cercignani, C. (1998). *Ludwig Boltzmann: the man who trusted atoms*. Oxford University Press.

Chalmers, D. J. (2006). Strong and weak emergence. In Clayton, P. and Davies, P., editors, *The Re-Emergence of Emergence*, pages 244–256. Oxford University Press, Oxford.

Clausius, R. (1864). *Abhanglungen über die mechanische Wärmetheorie*, volume 1. F. Vieweg und Sohn, Braunschweig.

Cooper, J. L. (1967). The foundations of thermodynamics. *Journal of Mathematical Analysis and Applications*, 17(1):172–193.

Crane, T. and Mellor, D. H. (1990). There is no question of physicalism. *Mind*, 99(394):185–206.

Curiel, E. (2014). Classical black holes are hot. arXiv: 1408.3691.

Dasgupta, S. (2014). The possibility of physicalism. *The Journal of Philosophy*, 111(9/10):557–592.

Davies, P. (1999). Is the flow of time an illusion? *International Framtider*, 9:4–8.

Davies, P. C. W. (1977). *The Physics of Time Asymmetry*. University of California Press, Berkeley.

Del Rio, L., Åberg, J., Renner, R., Dahlsten, O., and Vedral, V. (2011). The thermodynamic meaning of negative entropy. *Nature*, 474(7349):61.

Denbigh, K. and Denbigh, J. (1985). *Entropy in relation to incomplete knowledge*. Cambridge University Press, Cambridge.

Dennett, D. (2001). Are we explaining consciousness yet? *Cognition*, 79:221–237.

Dennett, D. C. (1991). Real patterns. *The Journal of Philosophy*, 88(1):27–51.

Dougherty, J. and Callender, C. (2016). Black hole thermodynamics: More than an analogy? http://philsci-archive.pitt.edu/13195/.

Duhem, P. M. M. (1902). *Termodynamique et chimie: leçons élémentaires*. A. Hermann & Fils, Paris.

Earman, J. (1974). An attempt to add a little direction to" the problem of the direction of time". *Philosophy of Science*, 41(1):15–47.

Earman, J. (2006). The ÂŞpast hypothesisÂŤ: not even false. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 37(3):399–430.

Earman, J. and Norton, J. D. (1998). Exorcist xiv: the wrath of maxwellÂŠs demon. part i. from maxwell to szilard. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 29(4):435–471.

Earman, J. and Norton, J. D. (1999). Exorcist xiv: The wrath of maxwellÂŠs demon. part ii. from szilard to landauer and beyond. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 30(1):1–40.

Eddington, A. S. (1928). *The Nature of the Physical World*. Ann Arbor Paperbacks.

Ehrenfest-Afanassjewa, T. (1925). Zur axiomatisierung des zweiten hauptsatzes der thermodynamik. *Zeitschrift für Physik*, 33(34):933–945.

Ehrenfest-Afanassjewa, T. (1956). *Die Grundlagen der Thermodynamik.* Leiden: E. J.Brill.

Einstein, A. (1919). Time, space, and gravitation (originally published in The London Times under the title: What is the theory of relativity?). In *Ideas and Opinions*, pages (pp. 227–232). Crown Publishers, New York.

Elkana, Y. (1974). *The Discovery of the Conservation of Energy*. Harvard University Press, Cambridge, Mass.

Elson, R., Hut, P., and Inagaki, S. (1987). Dynamical evolution of globular clusters. *Annual review of astronomy and astrophysics*, 25(1):565–601.

Feyerabend, P. (1966). On the possibility of a perpetuum mobile of the second kind. *Mind, Matter, and Method, eds PK Feyerabend, G. Maxwell, Univ. Minnesota P., Minneapolis*.

Feynman, R., Leighton, R., and Sands, M. (1964). *The Feynman lectures on physics. Vol. 2, Mainly electromagnetism and matter*. Addison-Wesley Pub., Reading, MA.

Feynman, R. P. (1996). *Feynman Lectures on Computation*. Penguin, London.

Fletcher, S. (2017). The principle of stability. manuscript.

Fodor, J. A. (1968). *Psychological explanation: An introduction to the philosophy of psychology*. Random House, New York.

Fodor, J. A. (1975). *The language of thought*, volume 5. Harvard University Press.

Frege, G. (1968). *The foundations of arithmetic: A logico-mathematical enquiry into the concept of number*. Northwestern University Press.

Frigg, R. (2010). A field guide to recent work on the foundations of statistical mechanics. In Rickles, D., editor, *The Ashgate companion to contemporary Philosophy of Physics*. Ashgate, Aldershot.

Frigg, R. and Hartmann, S. (2006). Models in science. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Gibbs, J. W. (1878). On the equilibrium of heterogeneous substances. *Trans. Connecticut Akad.*

Gibbs, J. W. (1903). *Elementary principles in statistical mechanics*. Dover (1960), New York.

Goldstein, S. (2001). Boltzmann's approach to statistical mechanics. In *Chance in physics*, pages 39–54. Springer.

Goodman, N. (1954). *Fact, Fiction, and Forecast*. University of London: Athlone Press, London.

Goold, J., Huber, M., Riera, A., del Rio, L., and Skrzypczyk, P. (2016). The role of quantum information in thermodynamics a topical review. *Journal of Physics A: Mathematical and Theoretical*, 49(14):143001.

Grünbaum, A. (1973). Is the coarse-grained entropy of classical statistical mechanics an anthropomorphism? In *Philosophical problems of space and time*, pages 646–665. Springer.

Hahn, E. L. (1950). Spin echoes. *Physical Review*, 80(4):580.

Halvorson, H. (2013). The semantic view, if plausible, is syntactic. *Philosophy of science*, 80(3):475–478.

Hare, R. M. (1952). *The language of morals*. Oxford Paperbacks.

Hawking, S. W. (1994). The no boundary condition and the arrow of time. *Physical origins of time asymmetry*, 37.

Heggie, D. (2003). The gravitational million-body problem. *Astrophysical Supercomputing Using Particles, IAU Symposium*, 208.

Heggie, D. and Hut, P. (2003). *The Gravitational Million-Body Problem*. Cambridge University Press, Cambridge.

Hellman, G. P. and Thompson, F. W. (1976). Physicalism: Ontology, determination, and reduction. *The Journal of Philosophy*, 72(17):551–564.

Hemmo, M. and Shenker, O. (2015). The emergence of macroscopic regularity. *Mind & Society*, 14(2):221–244.

Hempel, C. . P. O. (1948). Studies in the logic of explanation. *Philosophy of Science*, 15:135–175.

Horgan, T. (1993). From supervenience to superdupervenience: meeting the demands of a material world. *Mind*, 102.

Hornix, W. (1970). The laws of thermodynamics. *Pure and applied chemistry*, 22:535–538.

Horodecki, M. and Oppenheim, J. (2013a). Fundamental limitations for quantum and nanoscale thermodynamics. *Nature communications*, 4:2059.

Horodecki, M. and Oppenheim, J. (2013b). (Quantumness in the context of) resource theories. *International Journal of Modern Physics B*, 27.

Huang, K. (1987). *Statistical Mechanics*. Wiley, New York.

Hudetz, L. (2017). The semantic view of theories and higher-order languages. *Synthese*, pages 1–19.

Hut, P. (1997). Gravitational thermodynamics. *Complexity*, 3(1):38–45.

Janssen, M. (2009). Drawing the line between kinematics and dynamics in special relativity. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 40(1):26–52.

Jaynes, E. T. (1957). Information theory and statistical mechanics. *Physical review*, 106(4):620.

Jaynes, E. T. (1965). Gibbs vs Boltzmann entropies. *American Journal of Physics*, 33(5):391–398.

Joule, J. P. (1850). Iii. on the mechanical equivalent of heat. *Philosophical Transactions of the royal Society of London*, 140:61–82.

Kelvin, W. T. B., Larmor, J., and Joule, J. P. (1882). *Mathematical and Physical Papers: By Sir William Thomson...* At the University Press.

Kestin, J. (1979). *A course in thermodynamics*, volume 1. CRC Press.

Kiessling, M.-H. (1999). Statistical mechanics of gravitational condensation and the formation of galaxies. In *Galaxy Dynamics-A Rutgers Symposium*, volume 182, page 545.

Kim, J. (1998). Mind in a physical world: essays on the mind-body problem and mental causation. *J. Kim.–Cambridge, Massachusetts: MIT Press.–151 p.*

Kim, J. (1999). Making sense of emergence. *Philosophical studies*, 95(1-2):3–36.

Kim, J. (2002). The layered model: Metaphysical considerations. *Philosophical Explorations*, 5(1):2–20.

Kirchhoff, G. (1894). *Vorlesungen über die theorie der waerme*. Teubner, Leipzig.

Klein, M. (1967). Thermodynamics in Einstein's universe. *Science*, 157.

Knox, E. (2013). Effective spacetime geometry. *Studies in History and Philosophy of Modern Physics*, 3(44):346–356.

Knox, E. (2016). Abstraction and its limits: Finding space for novel explanation. *Noûs*, 50(1):41–60.

Kragh, H. (2001). *Quantum Generations: A History of Physics in the Twentieth Century.* Woodstock: Princeton University Press, Princeton, N.J.

Krylov, N. (1979). *Works on the Foundations of Statistical Physics*. Princeton University Press, Princeton.

Ladyman, J., Presnell, S., and Short, A. J. (2008). The use of the information-theoretic entropy in thermodynamics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 39(2):315–324.

Ladyman, J., Presnell, S., Short, A. J., and Groisman, B. (2007). The connection between logical and thermodynamic irreversibility. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 38(1):58–79.

Ladyman, J. and Robertson, K. (2013). Landauer defended: reply to norton. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 44(3):263–271.

Ladyman, J. and Ross, D. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press, Oxford.

Lam, V. Wüthrich, C. (2018). Spacetime is as spacetime does. arXiv:1803.04374v2.

Landau, L. and Lifshitz, E. (1969). *Statistical physics*. Pergamon, Oxford.

Landsberg, P. (1990). *Thermodynamics and statistical mechanics*. Dover.

Lavis, D. (2004). The spin-echo system reconsidered. *Foundations of Physics*, 34(4):669–688.

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.

Lavis, D. A. (2017). The problem of equilibrium processes in thermodynamics. http://philsci-archive.pitt.edu/12748/.

Lebowitz, J. L. (2007). From time-symmetric microscopic dynamics to time-asymmetric macroscopic behavior: An overview. *Boltzmann's Legacy*, pages 63–88.

Leff, H. and Rex, A. F. (2002). *Maxwell's Demon 2 Entropy, Classical and Quantum Information, Computing*. CRC Press.

Lévy-Leblond, J.-M. (1969). Nonsaturation of gravitational forces. *Journal of Mathematical Physics*, 10(5):806–812.

Lewis, D. K. (1970). How to define theoretical terms. *The Journal of Philosophy*, 67(13):427–446.

Lewis, D. K. (1972). Psychophysical and theoretical identifications. *Australasian Journal of Philosophy*, 50(3):249–258.

Lewis, D. K. (1986a). Causal explanation. In *Philosophical Papers vol. II*, pages p.214–240. Oxford University Press, Oxford.

Lewis, D. K. (1986b). *On the Plurality of Worlds*. Blackwell, Oxford.

Lewis, D. K. (1986c). A subjectivist's guide to objective chance. In *Philosophical Papers vol. II*, pages 83–133. OUP.

Lewis, D. K. (1987). *Philosophical Papers: Volume Il*. Oxford University Press, Oxford.

Lewis, D. K. (1988). Relevant implication. *Theoria*, 54(3):161–174.

Lieb, E. H. and Lebowitz, J. L. (1972). The constitution of matter: Existence of thermodynamics for systems composed of electrons and nuclei. In *Statistical Mechanics*, pages 17–24. Springer.

Lieb, E. H. and Yngvason, J. (1999). The physics and mathematics of the second law of thermodynamics. *Physics Reports*, 310(1):1–96.

Linden, N., Popescu, S., and Skrzypczyk, P. (2010). How small can thermal machines be? the smallest possible refrigerator. *Physical Review Letters*, 105(130401).

List, C. (2016). Levels: descriptive, explanatory, and ontological. http://philsci-archive.pitt.edu/12040/.

List, C. (2017). Levels: descriptive, explanatory, and ontological. *Noûs*, doi:10.1111/nous.12241(0).

List, C. and Pivato, M. (2015). Emergent chance. *Philosophical Review*, 124(1):119–152.

Liu, C. and Emch, G. (2002). *The logic of thermo-statistical physics*. Heidelberg: Springer.

Lloyd, S. (2006). Quantum thermodynamics: excuse our ignorance. *Nature physics*, 2:727–728.

Loewer, B. (2009). Why is there anything except physics? *Synthese*, 170(2):217–233.

Loewer, B. (2018). The mentaculus vision. In Loewer, B., Weslake, B., and Winsberg, E., editors, *Time's Arrows and the World's Probability Structure*. Harvard University Press.

Luczak, J. (2018). How many aims are we aiming at? *Analysis*, 78(2):244–254.

Lutz, S. (2017). What was the syntax-semantics debate in the philosophy of science about? *Philosophy and Phenomenological Research*, 95(2):319–352.

Lynden-Bell, D., Wood, R., and Royal, A. (1968). The gravo-thermal catastrophe in isothermal spheres and the onset of red-giant structure for stellar systems. *Monthly Notices of the Royal Astronomical Society*, 138(4):495–525.

Mainwood, P. (2006). *Is More Different? Emergent Properties in Physics*. PhD thesis, Oxford University., http://philsci-archive.pitt.edu/8339/.

Malament, D. B. and Zabell, S. L. (1980). Why Gibbs phase averages work–the role of ergodic theory. *Philosophy of Science*, 47(3):339–349.

Maroney, O. (2007). The physical basis of the gibbs-von neumann entropy. *arXiv preprint quant-ph/0701127*.

Maroney, O. (2009). Information processing and thermodynamic entropy. The Stanford Encyclopedia of Philosophy.

Masanes, L. and Oppenheim, J. (2017). A general derviation and quantification of the third law of thermodynamics. *Nature communications*.

Maudlin, T. (2017). (Information) Paradox Lost. https://arxiv.org/abs/1705.03541.

Maxwell, J. C. (1878). Diffusion. In *Encyclopedia Britannica (Nineth Ed.)*, volume 7, pages 214–221. Cambridge University Press.

Maxwell, J. C. (1891). *Theory of heat*. Longmans, Green.

McMullin, E. (1985). Galilean idealization. *Studies in History and Philosophy of Science Part A*, 16(3):247–273.

Mellor, D. H. (1978). In defense of dispositions. In *Dispositions*, pages 55–76. Springer.

Menon, T. and Callender, C. (2011). Turn and face the strange... ch-ch-changes: Philosophical questions raised by phase transitions. In Batterman, R., editor, *The Oxford Handbook of Philosophy of Physics*. Oxford University Press.

Milne, E. (1930). *Thermodynamics of the Stars*. Handbuch der Astrophysik. Springer, Berlin.

Moore, G. E. (1903). *Principia Ethica*, volume 960. Cambridge University Press.

Myrvold, W. C. (2011). Statistical mechanics and thermodynamics: A maxwellian view. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 42(4):237–243.

Myrvold, W. C. (2012). Deterministic laws and epistemic chances. In Yemima Ben-Menahem, M. H., editor, *Probability in Physics*, pages 73–85. Springer, New York.

Nagel, E. (1935). The logic of reduction in the sciences. *Erkenntnis*, 5:46–52.

Nagel, E. (1961). *The Structure of Science.* Harcourt, Brace and World, Inc., New York.

Ney, A. and Albert, D. Z., editors (2013). *The Wave Function: Essays in the Metaphysics of Quantum Mechanics.*, chapter 6, pages 52–57. Oxford University Press, New York.

Niven, W., editor (1965). *The Scientific Papers of James Clerk Maxwell*, volume 2. Dover Publications, New York, reprint of CUP edition of 1890 edition.

Norton, J. D. (2005). Eaters of the lotus: Landauer's principle and the return of maxwell's demon. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 36(2):375–411.

Norton, J. D. (2011). Waiting for landauer. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 42(3):184–198.

Norton, J. D. (2012). Approximation and idealization: Why the difference matters*. *Philosophy of Science*, 79(2):207–232.

Norton, J. D. (2014). Infinite idealizations. In *European Philosophy of Science–Philosophy of Science in Europe and the Viennese Heritage*, pages 197–210. Springer.

Norton, J. D. (2016). The impossible process: thermodynamic reversibility. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 55:43–61.

O'Connor, T. and Wong, H. Y. (2015). Emergent properties. In Zalta, E. N., editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University.

Oliver, A. (1996). The metaphysics of properties. *Mind*, 105(417):1–80.

Owens, D. (1989). Levels of explanation. *Mind*, 98(389):59–79.

Padmanabhan, T. (1990). Statistical mechanics of gravitating systems. *Physics Reports*, 188(5):285–362.

Papineau, D. (2010). Can any sciences be special. *Emergence in mind*, pages 179–197.

Penrose, O. (1979). Foundations of statistical mechanics. *Reports on Progress in Physics*, 42(12):1937–2006.

Phillips, A. C. (2013). *The physics of stars*. John Wiley & Sons.

Planck, M. (1926). Über die begründung des zweiten hauptsatzes der thermodynamik. *Sitzungsberichte der Preussischen Akademie der Wissenschaften*, 453–463.

Popper, K. R. (1957). Irreversibility; or, entropy since 1905. *The British Journal for the Philosophy of Science*, 8(30):151–155.

Poundstone, W. (2013). *The recursive universe: cosmic complexity and the limits of scientific knowledge*. Dover Publications, Mineola, New York.

Price, H. (1996). *Time's arrow and Archimedes' point: new directions for the physics of time*. Oxford University Press, Oxford.

Price, H. (2004). On the origins of the arrow of time: Why there is still a puzzle about the low-entropy past. In Hitchcock, C., editor, *Contemporary Debates in Philosophy of Science*, chapter 11, pages 219–239. Wiley-Blackwell, Oxford.

Price, H. (2009). The flow of time. In Callender, C., editor, *The Oxford Handbook of Philosophy of Time*. OUP, Oxford.

Prigogine, I. (1980). *From being to becoming: time and complexity in the physical sciences*. W. H. Freeman, San Francisco.

Prigogine, I. and Stengers, I. (1984). *Order out of chaos: Man's new dialogue with nature*. Flamingo, London.

Prunkl, C. (2018). The road to quantum thermodynamics. In C. Timpson, D. B., editor, *Quantum Foundations of Statistical Mechanics*. OUP, Oxford.

Prunkl, C. and Timpson, C. (2018). Black hole entropy is entropy (and not information). manuscript.

Putnam, H. (1967). Psychological predicates. *Art, mind, and religion*, 1:37–48.

Putnam, H. (1980). Philosophy and our mental life. *Readings in the Philosophy of Psychology*, 1:134–143.

Redhead, M. (1996). *From physics to metaphysics*. Cambridge University Press, Cambridge.

Reichenbach, H. (1991). *The direction of time*, volume 65. Univ of California Press.

Reif, F. (2009). *Fundamentals of statistical and thermal physics*. London : McGraw-Hill.

Ridderbos, K. (2002). The coarse-graining approach to statistical mechanics: how blissful is our ignorance? *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 33(1):65–77.

Ridderbos, T. and Redhead, M. (1998). The spin-echo experiments and the second law of thermodynamics. *Foundations of Physics*, 28(8):1237–1270.

Roberts, B. W. (2013). When we do (and do not) have a classical arrow of time. *Philosophy of Science*, 80(5):1112–1124.

Roberts, B. W. (2017). Three myths about time reversal in quantum theory. *Philosophy of Science*, 84(2):315–334.

Robinson, J. C. (2004). *An introduction to ordinary differential equations*. Cambridge University Press, Cambridge.

Rosaler, J. (2015). Local reduction in physics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 50:54–69.

Rosaler, J. (2017). Reduction as an a posteriori relation. *The British Journal for the Philosophy of Science*.

Rosenberg, A. (2008). *Darwinian reductionism: Or, how to stop worrying and love molecular biology*. University of Chicago Press.

Rowlinson, J. S. et al. (1993). Thermodynamics of inhomogeneous systems. *Pure and applied chemistry*, 65:873–873.

Rueger, A. (2006). Functional reduction and emergence in the physical sciences. *Synthese*, 151(3):335–346.

Ruelle, D. (1999). *Statistical mechanics: Rigorous results*. World Scientific.

Rumford, B. C. (1798). An inquiry concerning the source of the heat which is excited by friction. by benjamin count of rumford, frsmria. *Philosophical Transactions of the Royal Society of London*, pages 80–102.

Schaffner, K. (1967). Approaches to reduction. *Philosophy of Science*, 34:137–147.

Schlosshauer, M. A. (2007). *Decoherence: and the quantum-to-classical transition*. Springer Science & Business Media.

Sellars, W. (1963). Philosophy and the scientific image of man. *Science, perception and reality*, 2:35–78.

Shapiro, L. A. (2000). Multiple realizations. *The Journal of Philosophy*, 97(12):635–654.

Silberstein, M. (2002). Reduction, emergence and explanation. In *The Blackwell guide to the philosophy of science*, pages 80–107. Blackwell Oxford.

Sklar, L. (1993). *Physics and chance: Philosophical issues in the foundations of statistical mechanics*. Cambridge University Press, Cambridge.

Skyrms, B. (1977). Resiliency, propensities, and causal necessity. *The Journal of Philosophy*, 74(11):704–713.

Sober, E. (1999). The multiple realizability argument against reductionism. *Philosophy of science*, 66(4):542–564.

Spitzer, J. and Ostriker, L. P. (1997). *Dreams, stars, and electrons : Selected writings of Lyman Spitzer, Jr*. Princeton University Press.

Spitzer, L. (1987). *Dynamical evolution of globular clusters*. Princeton series in astrophysics. Princeton University Press., Princeton, N.J.

Strevens, M. (2008). *Depth: An account of scientific explanation*. Harvard University Press, Cambridge, Mass.

Szilard, L. (1929). Über die entropieverminderung in einem thermodynamischen system bei eingriffen intelligenter wesen. *Zeitschrift für Physik*, 53(11-12):840–856.

Teller, P. (1984). A poor man's guide to supervenience and determination 1. *The Southern Journal of Philosophy*, 22(S1):137–162.

Thompson, C. J. (1972). *Mathematical statistical mechanics*. Princeton University Press, New York.

Thomson-Jones, M. (2005). Idealization and abstraction: A framework. In Jones, M. and Cartwright, N., editors, *Correcting the Model: Idealization and Abstraction in the Sciences*. Rodopi, Amsterdam.

Tolman, R. C. (1938). *The principles of statistical mechanics*. Oxford University Press, Oxford.

Tong, D. (2012). Statistical physics. Cambridge Lecture Notes, available at http://www.damtp.cam.ac.uk/user/tong/statphys.html.

Uffink, J. (1996). Nought but molecules in motion. *Studies in History and Philosophy of Modern Physics*, 27(3):373–387.

Uffink, J. (2001). Bluff your way in the second law of thermodynamics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 32(3):305–394.

Uffink, J. (2006a). Compendium of the foundations of classical statistical physics, handbook for philosophy of physics, eds. *Butterfield, J. and Earman, J.*

Uffink, J. (2006b). Three concepts of irreversibility and three versions of the second law. In Stadler, F. and Stöltzner, M., editors, *Time and History Proceedings of the 28. International Ludwig Wittgenstein Symposium, Kirchberg am Wechsel, Austria 2005 (Publications of the Austrian Ludwig Wittgenstein Society – New Series (N.S.))*, pages 275–288. De Gruyter, Berlin.

Uffink, J. (2010). Irreversibility in stochastic dynamics. In Gerhard Ernst, A. H., editor, *Time, chance and reduction: philosophical aspects of statistical mechanics*, pages 180–207. Cambridge University Press, Cambridge.

Uffink, J. (2013). Three concepts of irreversibility and three versions of the second law. *From ontos verlag: Publications of the Austrian Ludwig Wittgenstein Society-New Series (Volumes 1-18)*, 1.

Valente, G. (2018). On the paradox of reversible processes in thermodynamics. *Synthese*.

Wald (1997). The "nernst theorem" and black hole thermodynamics. *Phys. Rev. D*, 56(10):6467–6474.

Wald, R. (2001). The thermodynamics of black holes. *Living Rev. Relativity*, 4(6).

Wallace, D. (2010). Gravity, entropy, and cosmology: In search of clarity. *The British Journal for the Philosophy of Science*, 61:513–540.

Wallace, D. (2011). The logic of the past hypothesis. http://philsci-archive.pitt.edu/8894/.

Wallace, D. (2012a). The arrow of time in physics. In Bardon, A. and Dyke, H., editors, *A Companion to the Philosophy of Time*. Wiley-Blackwell, Chichester, West Sussex ; Malden, MA.

Wallace, D. (2012b). *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford University Press.

Wallace, D. (2013a). Inferential vs. dynamical conceptions of physics. manuscript.

Wallace, D. (2013b). The non-problem of Gibbs vs. Boltzmann entropy. (Weblearn).

Wallace, D. (2014). Thermodynamics as control theory. *Entropy*, 16(2):699–725.

Wallace, D. (2015a). The quantitative content of statistical mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52:285–293.

Wallace, D. (2015b). The quantitative content of statistical mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 52:285–293.

Wallace, D. (2015c). Recurrence theorems: a unified account. *Journal of Mathematical Physics*, 56.

Wallace, D. (2016). Probability and irreversibility in modern statistical mechanics: Classical and quantum. In D. Bedingham, O. M. and (eds.), C. T., editors, *Quantum Foundations of Statistical Mechanics*. OUP.

Wallace, D. (2017). The case for black hole thermodynamics, part ii: statistical mechanics. *arXiv preprint arXiv:1710.02725*.

Wallace, D. (2018). The case for black hole thermodynamics part i: Phenomenological thermodynamics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*.

Weisberg, M. (2007). Three kinds of idealization. *The Journal of Philosophy*, 104(12):639–659.

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.

Wilson, J. (2009). Non-reductive physicalism and degrees of freedom. *The British Journal for the Philosophy of Science*, 61(2):279–311.

Wilson, M. (1985). What is this thing called "pain"?—the philosophy of science contemporary debate. *Pacific philosophical quarterly*, 66(3-4):227–267.

Woodward, J. (2005). *Making things happen: A theory of causal explanation*. Oxford University Press.

Zeh, H. D. (2007). *The physical basis of the direction of time*. Springer, Berlin, 5th edition.

Zwanzig, R. (1960). Ensemble method in the theory of irreversibility. *The Journal of Chemical Physics*, 33(5):1338–1341.

Zwanzig, R. (1961). Statistical mechanics of irreversibility. In *Lectures in Theoretical Physics (Boulder), Vol. 3*, pages 106–141. Inter-science, New York.