THE SCOPE OF THERMODYNAMICS

Carina E. A. Prunkl

Balliol College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hilary Term 2018

The Scope of Thermodynamics

Carina E. A. Prunkl

Balliol College
Doctor of Philosophy

Hilary Term 2018

Abstract

This thesis investigates the application of the laws of thermodynamics to various sub-systems of the universe. I begin by distinguishing between three different second laws: orthodox, statistical and probabilistic. After suggesting that entropy is best understood in means-relative terms, I then show that this interpretation does not imply an epistemic understanding of thermodynamics, pace Jaynes. I conclude the discussion of phenomenological thermodynamics by arguing that thermodynamics can properly be applied to systems containing only a single particle, given the right circumstances. Next I discuss thermodynamics in the context of black holes, and examine the question of whether black hole entropy is genuine thermodynamic entropy, as opposed merely to information-theoretic entropy. I examine the original arguments by Bekenstein and Hawking, and conclude that these are unsatisfactory, but I go on to demonstrate that black holes ought to be considered to be genuine thermodynamic objects by constructing a black hole Carnot cycle. The third chapter investigates thermodynamics in the quantum realm and begins by discussing a recent argument by Hemmo and Shenker against the identification of the von Neumann entropy with the thermodynamic entropy. Their argument is shown to be flawed as it a) allows for a violation of the second law and b) is based on an incorrect calculation of the entropy. I continue by providing a derivation of the laws of thermodynamics from quantum mechanics and a few phenomenological assumptions. This approach is then compared to so-called resource theories of thermodynamics and to single-shot thermodynamics. I end my discussion of quantum thermodynamics with the analysis of an argument made by Cabello et al., who claim that thermodynamics allows for the derivation of an empirical difference between two important classes of quantum interpretation. I provide a counterexample to this claim and show that the alleged heat cost is fully accounted for in the external agent.

Acknowledgements

After three and a half years of reading, thinking, writing and re-writing, I am now including this section in dedication to those who have supported me throughout this time.

I'd like to thank those who provided me with the financial means to complete my thesis: the British Society for the Philosophy of Science for their generous support throughout the first three years, the University of Oxford for the Vice Chancellor's grant that helped me weather through those last few months of thesis writing, Balliol College and the Templeton World Charity Foundation.

I am immensely grateful for the assistance I received from my supervisors. I would like to thank Chris Timpson for his steady support and his reassurance during the last few years. Without him and his encouragement, I would not be where I stand now. I am indepted to him also for showing me that academic prose need not be dry and grey but instead can be bonny and colourful. I would also like to thank Harvey Brown, my first philosophy of physics tutor who later became my second thesis supervisor. His habit of questioning even the most inocuous assumptions significantly contributed to my intellectual development (and lead to a lot of headscratching on my side). His guidance not only in academic but also in personal matters has been a main source of support during my thesis. I am also very grateful to Owen Maroney who guided me through the first, formative year of the DPhil and who provided me with countless insights into Statistical Mechanics. The conversations we had never failed to be intellectually stimulating and his views have had a large influence on my own. I would also like to express my gratitude to David Wallace, whose feedback on my work was invaluable, on point and always generously given. I am also very grateful for the useful comments and remarks of my examiner James Ladyman, that gave the thesis its finishing. Finally I would like to thank Adam Caulton, who as my College advisor always had an open ear for me, and Simon Saunders and Oliver Pooley, who all contributed to making Oxford the most wonderful place to study.

I would furthermore like to thank Erik Curiel and the Munich Center for Mathematical Philosophy for inviting me to Munich which gave me the opportunity to be part of a stimulating research community. I am equally grateful to Harvard's Black Hole

Initiative and in particular to Erik, Peter Galison and Paul Chesler for allowing me to significantly expand my horizon and to take part in many inspiring conversations during my stay in Cambridge. I would also like to thank the Centre for Quantum Technologies of the National University of Singapore for inviting me to discuss my work on Quantum Thermodynamics and Mile Gu for giving me valuable feedback.

Next I would like to mention my Oxford family, Tushar Menon, Niels Martens and Niels Linnemann, with whom I spent days and nights both inside and outside the cave. Their support, love and encouragement helped me through the rocky patches of thesis writing. I will always be grateful for this. I would in particular like to thank Tushar, my academic twin-brother, with whom I spent uncountable hours in Dosa Park to discuss work and life. I also want to express my gratitude to Katie Robertson, whom I might appoint as my personal motivator at some point in the future, as well as thank Franziska Poprawe, Laurenz Casser, Patrick Dürr, Andrea Ferrari, James Read, Julien Labonne and Julia Bird for their friendship and their advice.

I am also immensely lucky to have close friends in Berlin and Freiburg. The remaining Kleeeeblatt Lina Dreßler, Lena Teichler, Viktoria Heilmann, of course, but also Catrin Ciemer, Frank Gounelas, George Baroud and Thuan Bui were indispensable during these last few years and showed me that there was also a place outside Oxford that I could call home.

Finally I would like to thank my family, who never ceased to encourage me during the last few years. Their love and support carried me through good and difficult times and I would not be who I am today without them. In particular I would like to thank my partner Clément, who had to put up with me during the last few tough months of thesis writing and who always supported me but never complained.

Thank you.

Contents

Introduction

> Do not think that thermodynamics is only about steam engines: it is about almost everything.
>
> — Peter Atkins in *Four Laws that Drive the Universe*, Preface

Thermodynamics is a special theory indeed. A principle theory, a phenomenological one moreover, that outgrew its initial purpose of advancing nineteenth century engineering almost as soon as it was born[1]. Ceasing to be a mere tool for the optimisation of steam engines and the like, thermodynamics instead became a theory about, well, 'almost everything'. But this is not to say that thermodynamics can provide us with the answers to all scientific questions, quite the contrary: traditionally, it focusses on a narrow range of processes, singled out by the usage of terms such as heat, entropy and temperature. What makes thermodynamics so special, what sets it apart from other physical theories, is the comprehensive application of its laws *to* those other theories. Quantum information theory, quantum field theory, general relativity and even string theory, not to mention chemistry and biology—the laws of thermodynamics are ubiquitous among our scientific theories. And its influence seems to be growing still, with quantum thermodynamics developing into an ever so

[1] Kelvin as early as 1852 attributes the description of a 'universal tendency of natural processes' to thermodynamics, as Uffink (2001) points out.

active field of research, comprising quantum resource theories of thermodynamics and single shot quantum thermodynamics[2]. In the face of this long list, a suspicion arises: thermodynamics is no ordinary theory.

> 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 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 (1935), p.81 cited in Uffink (2001), p.5)

Eddington was not the only great thinker who recognised the extraordinary status of thermodynamics. Albert Einstein also famously admired the theory for its simplicity and extended range of applicability:

> [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. (Einstein, 1967, p.509)

Both the remarkable success of thermodynamics and the praise it has received by the scientific community become even more astonishing when we consider that there exists substantial ambiguity regarding the actual *content* of its laws, in particular the second law. To avoid confusion, I will call the unamended second law as originally formulated by Clausius and Kelvin (Clausius, 1864; Kelvin, 1882), the **orthodox second law**. In the Clausius version, it states that heat cannot pass from a colder to a warmer body without compensation[3]. However, with the rise of the kinetic theory of gases and the identification of temperature with average kinetic energy, it became

[2]Rio et al. (2011); Brandão et al. (2013); Horodecki and Oppenheim (2013); Skrzypczyk et al. (2014); Gour et al. (2015); Brandão et al. (2015); Goold et al. (2016).

[3]"Ein Wärmeübergang aus einem kälteren in einen wärmeren Körper kann nicht ohne Compensation stattfinden" (Clausius, 1887, p.82).

soon clear that due to molecular fluctuations, the second law is constantly violated on small sales. The velocities of the molecules of an ideal gas are not uniform, but instead distributed according to the Maxwell-Boltzmann distribution (Maxwell, 1860; Boltzmann, 1970). Due to random movement and constant collisions, the average kinetic energy of a subset of gas molecules may hence differ significantly from the average kinetic energy of a neighbouring subset, or the whole gas. This means, that within an ideal gas, there exist local temperature gradients. Such temperature gradients however constitute a violation of the orthodox second law, as heat is flowing sponatenously from a colder to a warmer body.

There are multiple ways to respond to this violation, resulting in different takes on how to understand the second law. Here are two options: i) concede that the orthodox second law is violated on small scales and take the second law to be a 'statistical truth' that holds for macroscopic systems, rather than a 'mathematical truth'. Or ii), amend the second law in such a way that microscopic fluctuations no longer violate it. The first option was advocated by James Clerk Maxwell, who took the second law to be true as long as the constituents of the thermodynamic system in questions are numerous enough and as long as we have no means of distinguishing between the individual molecules (Maxwell, 1871, p.308). With an increasing number of molecules in the system (or subsystem) under consideration, there will be less and less deviation from the overall mean, making a violation of the second law a 'practical impossibility' but not a mathematical one. I will call this version the **statistical version** of the second law.[4] Such a statistical understanding of the second law can still be found in some scientific and philosophical literature. Wüthrich (2017) for example writes that the second law is "not a strict law, but is only statistically

[4]See Myrvold (2011) for an extensive discussion of Maxwell's statistical understanding of the second law.

valid: it suffers from occasional, though very rare, violations" (p.4). Given that on a microscopic level the second law is not violated occasionally, but constantly, the last part of the sentence reveals that the scope of the second law is now restricted to systems of a certain size, namely macroscopic systems. But even for those systems, the second law is not strictly true, but only statistically.

Most scientists today however prefer the second option, ii), which re-interprets the *content* (as opposed to the scope) of the second law by giving it a probabilistic finish. The second law is then taken to forbid the *reliable* conversion of heat into work from a single reservoir by a cyclic process[5]. I will call this, in accordance with the literature, the **probabilistic version** of the second law (Ladyman et al., 2007; Maroney, 2009a; Myrvold, 2011). This is also the version used by Szilard (1929) and von Neumann (1996). The idea of reliability can be made mathematically precise by including restrictions on the expectation values of the relevant thermodynamic entities.[6] Formulated in terms of expectation values, the second law then holds strictly and can be given a rigorous statistical mechanical underpinning (see for example Penrose (1970)). The probabilistic version of the second law requires for its formulation the application of methods from statistical mechanics in order to make the notion of reliability precise. The use of probabilities in the context of thermodynamics however has been vehemently criticised by some philosophers and physicists (Goldstein, 2001; Albert, 2003).

Regrettably, the distinction between the orthodox, the statistical and the probabilistic second law is not always made sufficiently clear. This brings us to another large topic of discussion in the foundations of thermodynamics: the question of the meaning of

[5]In the statistical version, such a reliable conversion of heat into work is ruled out as a practical impossibility. In this probabilistic version, however, it is a strict impossibility.

[6]See Maroney (2009a) for an overview of the ambiguities associated with the concept of reliability in the context of the second law.

entropy. Von Neumann once famously admitted to Shannon that "nobody knows what entropy is" (von Neumann, quoted in Tribus and McIrvine (1971), p.179)[7] and the situation has not much improved since. Whereas in phenomenological thermodynamics, the orthodox second law defines the thermodynamic entropy, finding and interpreting the right statistical mechanical generalisation turns out to be tricky. If there even exists such a thing as *the* right generalisation, that is. In the classical case, the Gibbs and the Boltzmann entropies are the two most prominent candidates. However, each of them has their shortcomings[8], as will be discussed in the first chapter of this thesis. In the quantum case, the von Neumann entropy is the most promising candidate, as will also be discussed in length in Chapter 3.

Apart from the question of which entropy, there is also the question of how to interpret entropy. The answer will depend on which version of the second law one takes to define entropy and naturally also on which (if any) statistical mechanical generalisation one favours; and it will furthermore also depend on one's general outlook on thermodynamics, on what one takes thermodynamics to be *about*. Here is a selection of possible interpretations of entropy: one may consider it as being defined relative to the means of an observer, as Maxwell did (Maxwell, 1871) ; or as representing disorder, as some undergraduate textbooks claim (Giancoli, 2010; Nag, 2017); or as a measure of ignorance, as researchers working in quantum information theory, quantum thermodynamics but also black hole physics tend to do (Jaynes, 1957; Bekenstein, 1973; Lloyd, 2006; Rio et al., 2011); or as a property of the system's microstate, as supporters of a Boltzmannian statistical mechanics assert (Lebowitz,

[7]The full quote goes: "When Shannon first derived his famous formula for information, he asked von Neumann what he should call it and von Neumann replied "You should call it entropy for two reasons: first because that is what the formula is in statistical mechanises but second and more important, as nobody knows what entropy is, whenever you use the term you will always be at an advantage.""

[8]Ehrenfest and Ehrenfest (1909); Callender (1999); Goldstein (2001); Wallace (2013b); Albert (2003).

1993; Goldstein, 2001); as purely operational or as a property of the system. The list is long.

In this thesis, I will favour an operational understanding of thermodynamics, although very little of what is written here will actually depend on such a reading. Chapter 2 for example is completely neutral with respect to interpretations of entropy, or rather, it shows that a particular interpretation of entropy in information-theoretic terms is not needed for the case of black hole entropy. Chapter 3 will argue in favour of adapting the von Neumann entropy as the correct quantum statistical generalisation of the thermodynamic entropy[9], but will nevertheless remain neutral on how we ought to interpret this entropy.

The aim of this thesis instead is to find out what thermodynamics is *about* in the first place. Atkins asserts that thermodynamics is about 'almost everything', but what is 'almost everything'? When and to what objects do which laws of thermodynamics apply? There seems to be a substantial amount of disagreement in the literature about this. Lieb and Yngvason (1998) for example claim that a thermodynamic system "must be macroscopic, i.e, not too small". Furthermore, "systems that are too large are also ruled out because gravitational forces become important" (p.13). Such characterisations of thermodynamic systems however stand in direct opposition to current research practice in physics: quantum thermodynamics is all about deriving and analysing the thermodynamic behaviour of quantum systems, even individual, microscopic ones (Goold et al., 2016; del Rio et al., 2015; Horodecki and Oppenheim, 2013; Skrzypczyk et al., 2014). Likewise in the case of large gravitational objects: black hole thermodynamics is an ever so popular field in contemporary physics, focusing on the thermal behaviour of black holes (Bekenstein, 1973; Hawking, 1976; Wald, 2001). The importance of these issues has also been recognised by other

[9]By which I do not mean that the two should be considered identical!

philosophers of physics. Craig Callender for example asks:

> [W]hat sub-systems of the universe does [the second law] govern? Are the principles of thermodynamics responsible for generalizations about black holes? [...] What about the micro-realm? (Callender, 2016)

The principal aim of this thesis is to provide answers to these questions.

In the first chapter, I will introduce the laws of thermodynamics and furthermore discuss an operational interpretation of heat, work and entropy. It is based on Maxwell's understanding of entropy as a means-relative concept (Maxwell, 1871). This conception of thermodynamics as an agent-centric theory, whose primary concern is to determine what an external manipulator can or cannot do with a thermodynamic system, is also shared by Gibbs (1878) and Jaynes (1965). While I will discuss the advantages of such an interpretation of thermodynamics, I will also show that, *pace* Jaynes, adopting a means-relative understanding of thermodynamics does not imply anthropocentrism or subjectivism about entropy. I will then continue with a discussion of thermodynamics in the single particle regime. Thermodynamic systems containing only individual (or sufficiently few) particles are sometimes considered problematic, for the reasons mentioned above. I will argue that we may treat such systems as thermodynamic nonetheless, as long as they allow for a description in thermodynamic terms. The chapter will then proceed with an introduction of classical statistical mechanics and a discussion of the roles probabilities play in the Gibbsian statistical mechanical framework.

The second chapter begins the discussion about the scope of thermodynamics by investigating whether the laws of black hole mechanics simply are the laws of thermodynamics, but applied to systems containing to black holes. In short, I will examine whether black holes are fully fledged thermodynamic systems. In black

hole thermodynamics, geometrical properties of black holes are associated with classical thermodynamic variables. Since Hawking's discovery that black holes, when coupled to quantum matter fields, emit radiation at a temperature proportional to their surface gravity, the idea that black holes are genuine thermodynamic objects—with a well-defined thermodynamic entropy proportional to the black hole's horizon area—has become more and more popular. Surprisingly, arguments that justify this assumption are both sparse and rarely convincing. This has also been recently pointed out by Dougherty and Callender (2016) and Wüthrich (2017). The authors come to the conclusion that whatever black hole entropy may be, it certainly is not thermodynamic entropy. In Chapter 2 I will discuss and criticise the existing attempts to establish black hole horizon area (times a constant) as thermodynamic entropy. I will show that most arguments rest on an information-theoretic understanding of entropy, which, given the controversial status of information within the philosophical community, is undesirable. Despite the shortcomings of the pertinent arguments, I nevertheless maintain that black hole entropy is genuine thermodynamic entropy[10]. Furthermore, I will show how the criticism by Wüthrich (2017) against Bekenstein's original argument is largely unfounded. By means of a 'black hole Carnot cycle' I will then show how black holes, under certain conditions and given certain assumptions, do exhibit genuine thermodynamic behaviour. In particular, no information-theoretic notion of entropy is needed to derive this equivalence between black hole entropy and thermodynamic entropy.

In the second part of the thesis I will turn towards the theory of the very small: quantum mechanics. I will begin the chapter by recalling the original introduction of a quantum entropy by von Neumann (1996), which he made in his *Mathematische*

[10]This is in agreement with Wallace's recent analysis of black hole thermodynamics, in which he reaches the same conclusion. (Wallace, 2017)

Grundlagen der Quantenmechanik. This argument has recently been criticised by Shenker (1999) and Hemmo and Shenker (2006). I will show how their criticism is unfounded and in fact based on several misconceptions of the second law in the finite particle regime. But this still leaves open the question as to whether the von Neumann entropy really ought to be considered the correct quantum generalisation of the thermodynamic entropy, and the question of whether thermodynamics is applicable to the quantum regime in the first place. Based on work by Maroney (2007) and using well-known results in statistical mechanics[11], I will demonstrate how it is indeed possible to derive the thermodynamic behaviour of quantum systems by taking considerations from phenomenological thermodynamics and applying them to quantum mechanics. One of the main difficulties for any such enterprise is to show that the canonical distribution is indeed the correct quantum representation of thermal states. Maroney shows that assuming the existence of heat reservoirs, a concept well-known from phenomenological thermodynamics, allows one to derive this relation. Once the thermodynamic behaviour of quantum systems is established, the question of interpretation remains. Here it turns out that the interpretation of heat, work and entropy very much boils down to the interpretation of the quantum state itself, circumventing some of the criticism that is made against the use of probabilities in classical statistical mechanics. I will furthermore discuss a newly emerged field of research called quantum resource theories of thermodynamics, designed to derive the laws of thermodynamics from within a resource theoretic framework.

In Chapter 3, I will explore somewhat further the boundaries of thermodynamics by discussing a recent publication by Cabello et al. (2016), who claim that thermodynamics allows for the derivation of an empirically verifiable difference between two broad

[11] Gibbs (1878); Szilard (1925); Tolman (1938); Wehrl (1978); Partovi (1989); Gemmer et al. (2009).

classes of quantum interpretations. On the basis of three seemingly uncontentious assumptions, (i) the possibility of randomly selected measurements, (ii) the finiteness of a quantum system's memory, and (iii) the validity of Landauer's principle, and further, by applying computational mechanics to quantum processes, the authors arrive at the conclusion that some quantum interpretations (including central realist interpretations) are associated with an excess heat cost and are thereby untenable, or at least that they can be distinguished empirically from their competitors by measuring the heat produced. I will provide an explicit counterexample to this claim and demonstrate that their result can be traced back to a lack of distinction between system and external agent. It will be shown that the resulting heat cost in fact is fully accounted for in the external agent, thereby restoring the tenability of the quantum interpretations in question.

Chapter 1

Phenomenological Thermodynamics and Classical Statistical Mechanics

1.1 The Laws of Thermodynamics

Traditionally, thermodynamics is taken to be a phenomenological theory that is concerned with the behaviour of systems whose states are described by a small number of thermodynamic variables, such as temperature, volume, pressure and so on. Its core is built of five laws, which determine the possible transitions between these states. Central to thermodynamics is the notion of *equilibrium*, a state in which the state variables are non-changing in time. 'Phenomenological' means that the laws are empirical, based on experimental evidence, and that they are furthermore indifferent to the microscopic details of the system. I will use the expressions 'phenomenological thermodynamics' and 'traditional thermodynamics' interchangeably.

Minus First Law

The most fundamental of the laws of thermodynamics, and one that is often overlooked, is the minus first law (Brown and Uffink, 2001). It states that a system will spontaneously attain and remain in, unless subject to external change, a unique state of equilibrium, regardless of which state it was previously in. The time-asymmetry often associated with thermodynamic behaviour may be traced back to this fundamentally asymmetric law, as Brown and Uffink (2001) point out. The minus first law also implies that two systems that are brought into thermal contact equilibrate and reach a unique equilibrium state. The two systems are then said to be in *thermal equilibrium*.

Zeroth Law

The zeroth law links the notion of thermal equilibrium to the notion of temperature. It expresses the notion of 'being in thermal equilibrium' as an equivalence relation between systems, which in turn allows for the assignment of an *empirical temperature* to each system[1]. A consequence of the transitive nature of the zeroth law is that each of the systems is itself at equilibrium, and that locally the temperature is constant. It was Fowler and Guggenheim (1956) who coined the term 'zeroth law', but its most famous version is given by Planck (1897) as:

> If A is in thermal equilibrium with B, and B is in thermal equilibrium with C, then C will be in thermal equilibrium with A.

A different, axiomatic version of the zeroth law that emphasises the relationship

[1]The notion of empirical temperature is to be distinguished from the notion of *absolute temperature*, as introduced by the second law.

between thermal equilibrium and temperature is given by Sommerfeld: "Equality of temperature is a condition for thermal equilibrium between two systems or between two parts of a single system".[2]

The First Law

The meanings of *heat* and *work* in phenomenological thermodynamics emerge from the *first law of thermodynamics*,

$$dU = đQ - đW, \tag{1.1}$$

which states that the internal energy U of a closed system can only change by adding heat $đQ$ to the system and/or by performing work $đW$ on it. Work is the transfer of mechanical energy, for example the raising and lowering of a weight in a gravitational potential. Heat is then, by exclusion, a transfer of energy that does not lead to the raising and lowering of the weight.[3] The first law prohibits a *perpetuum mobile* of the first kind, i.e. a process that continuously extracts work from a system without an accompanying heat supply. Although energy conservation is readily found in other physical theories, the first law has its own subtleties: the inexact differentials in (1.1) indicate that heat and work are *not* functions of state, but solely particular kinds of *energy transfer*. A system therefore does not 'contain' heat or work. Instead, a *process* the system undergoes may be characterised by the amount of heat the system is exchanging and/or the amount of work performed on/by the system.

The first law, or indeed the thermodynamic laws in general, are often read as

[2] "Gleichheit der Temperatur ist eine notwendige Bedingung des thermodynamischen Gleichgewichts." (Sommerfeld, 1988, p.1)

[3] An alternative and more general distinction between work and heat defines work at 'usable' and heat as 'unusable' energy (Maxwell, 1878). Such a distinction relies on the specification of the *means* an agent has to manipulate the system, which I will discuss later in this chapter.

restricting the types of processes a thermodynamic system can undergo. The notion of a process, however, is still related to the system's state: a process is specified by an operation $\mathcal{O}$ on the thermodynamic system, and by a set of thermodynamic state variables $X_1, ..., X_n$. These variables parametrise the system's equilibrium states.

The Second Law

The first law provides us with a fundamental restriction on the class of allowable processes by demanding that they are energy-preserving, but it is ignorant about the *direction* of those processes. For this, we have the *second law of thermodynamics*. It imposes further constraints on which processes a system can undergo and focuses on the question of interconversion of heat and work. In the *Kelvin formulation*, the second law is stated as:

> 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 Version as given in (Uffink, 2001, p.26))

Special emphasis in this formulation is put on the fact that one is only considering *cyclic* processes and that this formulation of the second law mentions the existence of heat reservoirs. It will be shown later that heat reservoirs equally play a central role for the derivation of thermodynamic behaviour within a quantum mechanical setting. To give a precise definition of a heat reservoir (let alone a physical model for one) is non-trivial. Fermi (1956) defines it in the following way

> A body which is at the temperature [T] throughout and is conditioned in such a way that it can exchange heat but no work with its surroundings is called a [heat reservoir] of temperature [T]. ((Fermi, 1956, pp.29–30))

One may imagine such a body as a thermodynamic system enclosed in a rigid container, such that its volume can only change negligibly. A further desideratum is that the heat reservoir does not change its temperature upon the exchange of heat. And so, ideally a heat reservoir has an (approximately) infinite heat capacity, such that it remains (approximately) in the same thermodynamic state when it exchanges heat with other bodies. The question of how to represent heat reservoirs physically will be further discussed in Chapter 3.

Historically, the second law emerged from considerations about heat engines. Carnot in 1824 found that the efficiency ϵ of a heat engine operating in a reversible[4] cycle between two heat reservoirs with temperatures $T_1 < T_2$ is maximal and only depends on the temperatures of the reservoirs (Carnot, 1872) . This is also known as Carnot's theorem. Mathematically, the efficiency of a Carnot heat engine operating between two reservoirs can be written as

$$\epsilon = 1 - \frac{Q_1}{Q_2} = 1 - \frac{f(T_1)}{f(T_2)}, \tag{1.2}$$

where Q_1 is the heat absorbed by and Q_2 the heat ejected from the system. We can now define an absolute temperature scale $f(T) = T$ that is independent of the working substance used. Rewriting equation (1.2) and adapting the signs of the heat term Q_i such that it is positive if heat is absorbed and negative if heat is ejected, results in

[4]The usage of the term 'reversible' bears some ambiguity, as Uffink (2001) points out. 'Reversible' may refer either to the ability to ' undo' a process such that the original state of the system is recovered. Or it may refer to the ability to run the process infinitely slowly (quasi-statically), in which case the process can also be performed 'backwards' (although Uffink denies this conclusion). The literature and its translations are quite heterogeneous in this respect, with Kelvin and Planck taking ' reversible', or 'reversibel' in Planck's case, to mean 'recoverable', whereas Fermi and Clausius, for example, take it to refer to 'quasi-static' processes. I will try to use 'reversible' and 'quasi-static', to make the distinction clear.

$$\sum_i \frac{Q_i}{T_i} = 0. \quad \text{(reversible cycle)} \tag{1.3}$$

For an infinitesimal amount of heat $đQ$ absorbed, the above equation becomes:

$$\oint \frac{đQ}{T} = 0 \quad \text{(reversible cycle)}, \tag{1.4}$$

where the integral is taken over a quasi-static path, i.e. the process takes place infinitely slowly, such that at every instance the system is found at equilibrium. The temperature $T > 0$ notably refers to the temperature of the heat reservoirs. If the system is in thermal equilibrium with the reservoir, the zeroth law ensures that T locally is also the temperature of the system.

Under the application of the first and second law, it is now possible to show that the above equation changes into the famous Clausius inequality, with equality only for reversible cycles:

$$\oint \frac{đQ}{T} \leq 0. \quad \text{(Clausius inequality)} \tag{1.5}$$

To derive the entropy function we now take a reversible cycle to consist of two quasi-static transformations between two equilibrium states s_i and s_f. Then equation (1.4) becomes

$$\oint \frac{đQ}{T} = \int_1^2 \frac{đQ}{T} + \int_2^1 \frac{đQ}{T} = 0. \text{ (reversible cycle)} \tag{1.6}$$

The above equation allows us to define a state function, the *thermodynamic entropy*. It is given by

$$\Delta S = S_f - S_i = \int_{s_i}^{s_f} \frac{đQ}{T}, \text{ (reversible)} \tag{1.7}$$

where the integral is taken over a quasi-static path. In phenomenological thermodynamics, an entropy function is therefore only defined for equilibrium states. It is furthermore defined only up to an arbitrary additive constant. As a state function, the entropy of a system only depends on the current state of the system.

For general transformations, equation 1.7 becomes an inequality

$$\Delta S = S_f - S_i \geq \int_{s_i}^{s_f} \frac{đQ}{T}, \tag{1.8}$$

where the integral again is taken over a quasi-static path. For isolated systems (systems that cannot exchange any heat or matter with the environment), $đQ = 0$, and so the above inequality becomes

$$S_f \geq S_i. \tag{1.9}$$

This is the *Entropy-version* of the second law. In words, it states that

> [F]or any transformation occuring in an isolated system, the entropy of the final state can never be less than the entropy of the initial state. (Fermi, 1956, p.55)

The entropy of a system can change either by internal entropy production, such as the free expansion of a gas, or by a heat exchange with another system or the environment.

We can re-formulate the second law for *isothermal* processes, i.e. processes during which the system is kept at a constant temperature throughout, in terms of the Helmholtz free energy F:

$$F = U - TS \tag{1.10}$$

which including the above restrictions on processes amounts to

$$đW \geq dF \tag{1.11}$$

where W is the work performed on the system and dF the change in Helmholtz free energy. Equality is reached for isothermal, reversible processes. The free energy therefore gives us a limit for the maximally extractable work of a thermodynamic system in contact with a heat bath.

In this section I introduced phenomenological thermodynamics by following the arguments of Carnot, Clausius, Kelvin and Planck (Carnot, 1872; Clausius, 1887; Kelvin, 1882; Planck, 1897), but there naturally exist other accounts of thermodynamics. Carathéodory and Lieb and Yngvason for example follow a different route (Carathéodory, 1909; Lieb and Yngvason, 1998). Within their axiomatic frameworks, entropy is defined in terms of 'adiabatic accessibility'. In this case, no explicit reference to cycles needs to be made in order to derive the second law. Even though structurally, there are striking similarities between such axiomatic approaches and the below discussed 'quantum resource theories of thermodynamics' (Section 3.2.6), I will forgo a thorough account of this framework due to lack of space.

1.2 Interpreting the Phenomenological Entropy

The laws of thermodynamics as introduced in the previous section describe the way thermodynamic systems change their state when subject to certain processes. A

system's state is given by a set of state variables $X_1, ..., X_n$, such as energy and volume, and a thermodynamic system is a system that allows for such description. Entropy is such a state variable, which is why entropy, as opposed to heat and work, is part of the thermodynamic state of a system. For this reason, it is often said that entropy is a *property* of the thermodynamic system in question. This becomes especially important in the context of finding a statistical mechanical generalisation of the thermodynamic entropy. It will be shown later in this chapter that one of the main objections to the so-called Gibbsian account of statistical mechanics is that the Gibbsian framework supposedly commits us to an epistemic reading of entropy.[5] The prospect of entropy being a mind-dependent entity, however, is not embraced by everyone:

> A steam engine's efficiency doesn't care about whether anyone is looking, our uncertainty, or our beliefs. (Dougherty and Callender, 2016, p.21)

Prima facie this is right: whether or not a locomotive is running better be a non-epistemic matter of fact that depends on the construction, and operation, of the engine, but not on whether anyone is looking. And yet, there nevertheless does exist an element in thermodynamics that very much seems to relate to an agent's *ability*, as I will now discuss. Bridgman (1941) for example remarks that thermodynamics "smells more of its human origin than other branches of physics" (p.214), because the first and second law are normally understood as describing what an implicitly present manipulator can or cannot do with the thermodynamic system in question. This reference to the manipulations an agent can or cannot do seems suspiciously

[5] I will distinguish between non-epistemic and objective. In the first case I am referring to the property of being mind-independent, that is independent of knowledge or perspective. 'Objective' describes an inter-subjective agreement between agents, as one for example encounters in objective Bayesian approaches (see Denbigh and Denbigh (1985) for a distinction between these two types of objectivity.)

different from the rest of our physical theories. The phase space trajectory of a bouncing ball, the worldline of a particle, makes no reference to which manipulations and interventions can be performed on the system. We can trace back the source of Bridgman's comment on the 'human smell' of thermodynamics, to his understanding of the term 'possibility': the standard formulations of the second (and sometimes the first) law are all phrased in terms of 'possible' and 'impossible' processes. The second law as given above states that it is *impossible* to perform a cyclic process, if the only result is that heat is absorbed from a reservoir and work is performed. Whether or not this impossibility is a nomological possibility, i.e. something forbidden by the laws of nature, or a practical impossibility, i.e. something that is not ruled out by our physics, but could in principle be done if one had the right expertise, technology, resources, and so on, is not obvious from the way the laws are stated.

Given that the orthodox second law (as given by Kelvin, Planck and Clausius) is *de facto* violated on a regular basis due to fluctuation phenomena, it becomes clear that 'possibility' can only refer to practical possibility[6]. And suddenly we are, conceptually at least, not very far from talk about beliefs: if the second law merely refers to a practical impossibility, then two agents with different means will disagree on what the entropy of a given system is. Such means-relative understanding of the thermodynamic entropy[7] at least superficially lends itself to justifying an anthropocentric account of entropy, as long as 'means' are associated with states of knowledge.

In this section, I will explore this idea further, in order to determine a) whether the fact that the orthodox second law can only refer to a practical impossibility forces

[6]This only refers to the orthodox second law and not to the probabilistic version. I will come back to this point shortly.

[7]As far as I know the term 'means-relative' was coined by Myrvold (2011), who takes this to be Maxwell's interpretation of entropy and the second law.

us to take on an anthropocentric notion of entropy and b) what the consequences of a means-relative interpretation for our understanding of entropy are.

Maxwell, Jaynes and the Gibbs Paradox

According to Myrvold (2011), Maxwell was the first to develop what was labelled a means-relative approach to entropy and the second law. Other prominent physicists, such as Gibbs (1878) and Jaynes (1965) promoted very similar views[8] and the control-theoretic understanding of thermodynamics by Wallace (2014) falls into the same category.

To motivate a means-relative understanding of entropy, I will first consider the familiar Gibbs paradox (Gibbs, 1878), which is elegantly resolved by this approach. We first consider a box that contains two monoatomic gases, separated by a partition, but both at the same temperature and pressure. The partition is removed and after the mixture has settled into equilibrium again, it is re-inserted into the box. The paradox is that depending on whether the molecules belonging to the two gases differ in the slightest or are the same, the entropy behaves differently during the process. If the two gases are of the same kind, thermodynamics predicts that the entropy remains constant throughout the whole process. But if the two gases differ, the entropy of the total system increases by an amount $\Delta S = nR \ln 2$. This is often called the entropy of mixing. This entropy increase is predicted by thermodynamics, even if the two gases share all the *relevant* properties, but differ only in some unimportant way. By relevant properties I mean properties that are needed for the calculation of the molecules' trajectories or for the prediction of their behaviour when they come in contact with the other gas' molecules, like mass, charge and so on. The paradoxical

[8] I am here concerned with interpretations of the phenomenological second law. Jaynes very famously also defended an unambiguously anthropocentric view of the statistical mechanical entropy.

situation can be made vivid by the following: even if the only difference between the molecules in the gases is that molecules X_1 react in a funny way with some unknown extraterrestrial element Y, whereas molecules X_2 do not, even then thermodynamics predicts an entropy increase when the two gases mix. The molecules of the two gases may even follow the same *trajectories* — thermodynamics predicts an increase in entropy. This surprising dependence of thermodynamics on microscopic details of a gas, that may not even play any role in our calculation for the molecules' behaviour, is called the Gibbs paradox.

Understanding heat, work and entropy as means-relative, however, resolves the Gibbs paradox[9]. Historically, the means-relative understanding of thermodynamics has its origin in the tension between phenomenological thermodynamics and the kinetic theory of gases: in the kinetic theory of gases, temperature is proportional to the average kinetic energy of the molecules in the gas. A being with the capacity of tracking and manipulating the individual molecules could in principle extract all of the gas' kinetic energy as work, thereby violating the second law[10]. Famously, such a being is called a Maxwell demon, after its inventor. Although it is now widely agreed upon that an amended version of the second law cannot be violated by such a demon, Maxwell's thought-experiment nevertheless demonstrates that the difference between work and heat cannot be fundamental: a demon who can track and manipulate molecules at its will may as well be unaware of the existence of what we normally call heat. What is work and what is heat instead depends on the means available for the manipulation of the thermodynamic system. Maxwell emphasises this observer-dependence of heat and work by dividing energy into 'available' and 'dissipated' energy:

[9]Note that there are various ways of resolving the Gibbs paradox. A thorough account that is based on the indistinguishability of particles is given for example by Saunders (2018).

[10]If the demon is not modelled as a physical system, it could reliably violate the second law.

> 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. (Maxwell, quoted in (Myrvold, 2011, p.7))

Finally, if heat and work are taken to be means-relative, entropy must equally be understood as relative to the means an agent has to manipulate the thermodynamic system in question. Once entropy is considered as defined with respect to those means, the Gibbs paradox ceases to be paradoxical. Here is why (Jaynes, 1965): in a means-relative approach, entropy is understood as being relative to the means an agent has to analyse and manipulate the system. For an agent that has *not* the relevant means to separate the gases again, their mixing is something like a lost opportunity to extract work from the system. Hence the entropy increases. If on the other hand an agent *knows*, or rather, *has the means* to distinguish between the two gases, she can in principle use this knowledge to extract work from the system, for example by inserting specifically designed semi-permeable membranes that allow for the separation of the gas. In this case the system could be returned into its original state.

What does it mean for the gases to be brought back to their original state? It certainly does not mean that each of the molecules returns to its original position. It means that the system returns to its original *thermodynamic* state, described by the set of macrovariables $X_1, ..., X_n$. Entropy is a function of this set of macrovariables $S = S(X_1, ..., X_n)$[11]. If the variables return to their initial value, we say that the system has returned to its initial state. In the case of distinguishable gases, this is

[11] In light of the repeated claim that the statistical mechanical Boltzmann entropy refers to a property of a system's microstate: note that here entropy, as a function of the macrovariables, is in fact a property of the thermodynamic macrostate. It therefore does not refer to a property of a particular microstate, but rather to a reference class of microstates.

clearly not the case after the partition has been re-inserted.

In the case of the two gases being the same on the other hand, thermodynamics predicts that the thermodynamic state of the system before and after the partition is removed *is* the same, even though the individual molecules will have mixed during the time the partition was out. And so, according to Maxwell, to say that two gases are (thermodynamically) the same is just to say that one "cannot distinguish the one from the other *by any known reaction*." ((Maxwell, 1878, p.221, emphasis added) quoted in (Myrvold, 2011, p.8)). Not only does Maxwell make reference to 'knowledge', his statement also implies that there may exist yet undiscovered ways of distinguishing between the two gases after all. Saying that the two gases are (thermodynamically) different, is just to say that one *can* separate them again, if one has the means to analyse and manipulate the gases.

A further consequence of the means-relative approach is that two different agents might have access to different means (Maxwell, 1871; Jaynes, 1965; Myrvold, 2011). If Alice, for example, has some special powers to distinguish between two gases but Bob takes them to be the same, then the two will assign different states to the thermodynamic system in question. Alice takes the thermodynamic state of the system to be given by $X_1, ..., X_{n+1}$, whereas Bob only has access to variables $X_1, ..., X_n$. In this particular case, Alice could insert semi-permeable membranes and extract work from the system, whereas Bob couldn't do any such thing. Even worse, for Bob it would seem as if Alice was violating the second law. From the point of view of Alice, however, no violation of the second law has taken place whatsoever, as the thermodynamic state of the system before and after her manipulation is a different one. More importantly, whilst the available processes an agent can perform on the system are relative to the chosen variables (and hence the means), whether

the second law is actually violated is *not* relative to anyone's perspective. This means that even though it *seems* to Bob that Alice was violating the second law, in fact, she is not.

Means and Anthropocentrism

Agents assigning different entropies to the same system brings to light the sense in which thermodynamics 'smells of its human origins'. The question is, whether the means-relative approach has revealed the existence of an anthropocentric view of thermodynamics, one in which entropy is fundamentally a property of an agent's beliefs, as opposed to a property of the thermodynamic system in question.

No relief is found by turning to the main advocates of the means-relative approach. Maxwell writes about the anthropocentric nature of the heat/work distinction:

> [...] 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. (Maxwell (1878), reprinted in (Maxwell, 1965) and quoted in Myrvold (2011))

Myrvold (2011), in his analysis of Maxwell's writing, equally asserts that the distinction "seems to rest on anthropocentric considerations" (p.6), which he takes to be the same as being considerations of the available means of an agent. Jaynes (1965), too, takes the step towards anthropocentrism. But he goes even further, denying even the *physicality* of entropy[12]:

> The strong contrast between the 'physical' nature of energy and the 'anthropomorphic' nature of entropy was well understood by Gibbs before 1875.(Jaynes, 1965, p.6)

[12]Physicality in this context can be understood as the property of being mind-independent. As such, physicality it stands in contrast with what Jaynes considers an anthropomorphic entropy.

Considering that Jaynes is famously known for his epistemic account of Gibbsian statistical mechanics, it may not be too surprising that he is equally prone to taking the *thermodynamic* entropy as being 'anthropomorphic'. But this jump from means-relative to anthropocentric is too quick. The rest of this section will discuss why thermodynamics, even if understood as a means-relative theory, is not worryingly anthropocentric. We can divide the problem by taking a closer look at what it means to be means-relative. The first component is that entropy so understood is defined *relative* to a set of parameters. The second component is that which set of parameters is chosen to be representative of the system, may depend on an external agent's state of knowledge.

First: Relative Quantities

Granted, entropy is understood as relative to a set of variables. But this alone is no good reason to claim that entropy is anthropocentric. There exist many notions in physics that are defined as relative to other things. *Equilibrium* is an example taken directly from thermodynamics. To say that a system is at equilibrium, means to say that the system is at equilibrium relative to a given set of variables, meaning these variables are non-changing in time, which is usually implicit in the setup. But this in no way implies that equilibrium is anthropocentric. Another example is *energy*. In special relativity, kinetic energy is reference frame dependent. A body at rest in one reference frame will have zero kinetic energy. The same body will have positive kinetic energy when analysed from within a moving reference frame. And so, despite the fact that kinetic energy only really is a useful quantity when defined relative to a given reference frame, this relativity certainly does not imply that kinetic energy must be considered anthropocentric. Relativity alone is therefore not a sufficient

criterion for the claim that entropy is anthropocentric.

Second: Choice of Variables

A second worry one may have is that the choice of variables relative to which entropy is defined seems to be determined by our knowledge of the system. In other words, if the choice of variables is means-relative, that is: the choice of variables depends on our means, which in turn are determined by our knowledge of the system, then means-relative seems to collapse into mind-relative. But it need not. Instead, we can easily consider the choice of variables to refer to a choice of description. There is a priori nothing subjective about preferring one set of variables over another to describe a system, something that is done all the time in science. To describe the movement of an ant, for example, it would be whimsical to track the movement of all of the ant's cells or molecules. Instead we chose a coarse grained description of the world, we split it into 'ant' and 'environment' and use the position of the coarse-grained ant as a variable. This allows us to better represent the relevant physical quantities. Choosing a coarse-grained description of the world thereby has nothing a priori anthropocentric about it. Equally, whether this coarse-grained description is out of free choice (because it's useful) or out of necessity (because I have no idea about what the molecules of an ant are doing) does not seem to weaken this claim. One thing that is clear is that usually in science, coarse-grained variables are used because they are more suited for the purpose in question. Entropy is not different from variables in the rest of science in this regard.

Coming back to thermodynamics: the distinction between heat and work may be a natural distinction to draw at a certain level of abstraction, but not at a more fine-grained level (the level a demon would be operating on). Furthermore, to describe a

system by a set of coarse-grained variables $X_1, ..., X_n$ does not force us to understand these variables as having anything to do with lack of knowledge. The fact that two agents chose different levels of descriptions for the same system, and can therefore perform different sets of operations (which are always taken to be functions of the set of variables in question), is therefore more a matter of description, rather than a matter of mind.

To summarise the above: there is nothing a priori subjective or epistemic about a means-relative understanding of thermodynamics, in which entropy is defined relative to a set of macroscopic variables. The importance of a means-relative understanding of thermodynamics will become evident when the so-called quantum resource theories of thermodynamics are discussed in Chapter 3. Structurally very similar to the means-relative approach, these theories show in an impressively simple way how thermodynamic-like behaviour arises from dividing the set of states into resources and free states and determining which transitions are possible given a set of allowed operations. Unlike the means-relative understanding of thermodynamics discussed above, that focuses only on phenomenological thermodynamics, quantum resource theories focus on a quantum mechanical generalisation of thermodynamics.

Another extension to the means-relative approach described above is Wallace's take on thermodynamics as a *control-theory* (Wallace, 2014). A control-theory studies the behaviour of systems when exposed to a range of external interventions, so-called control operations. An advantage of such an approach is that feedback-processes are easily understood in this framework.

In the next section, I will continue with the discussion of phenomenological thermodynamics by exploring whether the notions of heat, work and entropy are well-defined for systems containing only a single particle. This will be of importance for the dis-

cussion of the argument against the von Neumann entropy by Hemmo and Shenker (2006) in Chapter 3.1, as their argument is based on the thermodynamics of a one-molecule system.

1.3 Phenomenological Thermodynamics for Individual Particles

Prima facie, we observe a tension within the literature: on the one hand, thermodynamics is said to be applicable only to macroscopic systems (Lieb and Yngvason, 1998) and the second law only reliable for systems containing large numbers of molecules such that the likelihood of observing statistical fluctuations becomes negligibly small[13]. On the other hand, one-molecule systems have become a prime example for exorcising Maxwellian demons or proving the non-equivalence of thermodynamic and von Neumann entropy (Szilard, 1929; Hemmo and Shenker, 2006). In each case, phenomenological thermodynamics is applied to single particles without much further justification.

I will first discuss the origin of this tension, before showing that thermodynamics can be safely applied to the one-particle as well as to the many-particle realm. In the introduction, I distinguished between three second laws: orthodox, statistical and probabilistic. Maxwell adopted a statistical stance, because he observed the violation of the orthodox second law by fluctuation phenomena on small scales. If the validity of the second law decreases with the number of particles, then it seems reasonable to assume that traditional thermodynamics cannot apply to systems containing only a single particle. But, digging deeper, we can ask what we mean by a single particle

[13]Norton (2011) even explicitly argues against the use of one-molecule systems in thermodynamic arguments.

system in the first place: does the term refer to one particle, maybe among others, maybe alone, moving around freely in the world; or does the term refer to a particle in a box with perfectly reflecting walls, maybe even in contact with a heat bath. In the former case, we are concerned with a mechanical system, in the latter case with a thermodynamic system. The difference again boils down to the level of description that is chosen to characterise the system in question. A thermodynamic system may be described in mechanical terms, but *thermodynamics* only applies to systems describable by a set of conserved parameters and external constraints. And a single particle enclosed in a box with volume V can be described in thermodynamic terms.

When I speak to colleagues about single particle thermodynamics, the first response I normally get is: "But what about equilibrium?" And so, indeed, what about equilibrium? As a requirement for the definition of the thermodynamic entropy, the notion of equilibrium is certainly central, and the idea of a single particle being at equilibrium seems counter-intuitive in many ways. This is not least due to the image one may have in mind from statistical mechanical textbooks, that associate equilibrium with the spread of molecules within a volume. Furthermore, as will be discussed in the next part of this chapter, in a Boltzmannian statistical framework, equilibrium corresponds to the largest macrostate the system can be in, and it is unclear what such an equilibrium macrostate should correspond to, in the case of an individual particle.

But these worries are all based on statistical mechanics, not on thermodynamics. In thermodynamics, the system can be considered a black box, and there is no mentioning of molecules or how they are distributed within the box. Furthermore, we may consider two different types of equilibrium the thermodynamic system can be in: internal and external. Internal equilibrium corresponds to the property of an

isolated thermodynamic system to be parametrisable by a set of thermodynamic variables that are non-changing in time[14]. External equilibrium, on the other hand, corresponds to a relational property between two systems, to being in 'equilibrium with'. It means that there is no net-flow of heat, energy or particles between the two systems. A single particle at equilibrium with a heat bath, say, is hence a well defined notion[15]. Worth mentioning is also that all notions of equilibrium are always understood relative to a given set of variables.

With this external notion of equilibrium in mind, we can now explain why it is meaningful to assign an entropy to the one-particle system. In the original derivation of the second law, the temperature appearing in the Clausius inequality corresponds to the temperature of the heat bath[16]. The notion of thermodynamic entropy thus depends first and foremost on the system's ability to be at external equilibrium with a heat bath at temperature T, which is the case for a single particle in a box. Therefore, it is meaningful to assign a thermodynamic entropy to such systems. Furthermore, we can do most standard thermodynamic operations on a box containing a single particle. For example, it is possible to expand the box quasi-statically[17] by having the particle 'push' a piston, thereby extracting work from the system. It is also possible to do this expansion isothermally by coupling it to a heat bath at temperature T.

Whereas some operations such as free or isothermal expansions or compressions are unproblematic for single-particle systems, others, such as the division of a one-particle system by inserting a partition into the middle of the box, are less innocent,

[14]Brown and Uffink's 2001 minus first law corresponds to the statement that thermodynamic systems approach and remain in such equilibrium states.

[15]I would even concede to the point that heat baths are necessary to make sense of a single particle system.

[16]Tolman (1938) for example is explicit about this fact and so is Maroney (2007).

[17]Opponents might now object that a expansion of the box requires the particle to bounce off the piston many times and that this requires the dynamics of the particle to be of a special kind. But this is just a quantitative, not a qualitative difference between a single-particle gas and a many-particle gas, which also needs to 'push' the piston by bouncing off it.

because here work can be extracted from the system by attaching a tiny weight to the partition and letting the molecule isothermally expand. I very much agree with von Neumann, who declares any thermodynamic process as valid provided that it is not in disagreement with the laws of thermodynamics (von Neumann, 1996, p.192). But here we may rightly ask: *which* laws? Certainly the orthodox law is violated in the above (in agreement with the statistical conception described in the introduction). The question of which laws *seems* to become particularly urgent in the case of one-molecule systems, but this can be explained by the fact that fluctuations are much easier exploited when the system contains fewer particles. In the case of one-molecule systems, the exploitations are especially dramatic, and so I will stick with discussing one-molecule systems, even though the discussion generalises to any finite-particle system. The most famous one-particle system is Szilard's engine, which will play a role for the argument by Hemmo and Shenker which I discuss in Chapter 3.1.

Szilard's Engine and Maxwell's Demon

Szilard's 1929 *Gedankenexperiment* goes as follows: we consider a single molecule in a box in contact with a heat bath. At some point, a partition is inserted into the middle of the box. The molecule is then allowed to expand isothermally, thereby lifting a tiny weight that is attached to the partition. The process is then repeated as often as desired, thereby reliably extracting heat from a heat bath and converting it into work — a violation of the second law (both strict and probabilistic).

The exorcism typically put forward is that in order to attach the weight, we (or the demon) need to *know* on which side of the box the molecule is found (put differently and without reference to knowledge, it is impossible to design a machine that would

reliably lift a weight no matter which side the molecule would be trapped in). If we model the demon as a physical system, it therefore first needs to perform a *measurement* on the system, which leads to a change in the demon's own state. The simplest version of such a demon is a so-called *memory cell*, which again consists of a molecule in a box with two compartments that align with the system-molecule's position after a measurement. At the end of the cycle, it is then required that *both* system and demon return to their original state.

Originally it was thought that there existed an entropy cost associated with the *measurement* of the particle's position (Smoluchowski, 1913; Szilard, 1929), balancing out the apparent violation of the second law, until Bennett (1973) showed that measurements could be performed without dissipation. It is now, at least among physicists[18], widely accepted that the 'missing' entropy is recovered by resetting the demon's memory, i.e. by bringing the demon back into its original state (Bennett, 1973; Landauer, 1961), which cannot be implemented without a heat transfer into the environment — an insight nowadays known as *Landauer's principle*. Landauer's principle recovers the second law if the total entropy of both system and and memory cell are taken into account.[19]

The lesson to be drawn from the Szilard Gedankenexperiment is that whenever an agent performs an operation whose particular execution depends on the outcome of a measurement, the entropy contribution of the measurement apparatus, including its resetting, in a cyclic process, needs to be taken into account.

Let me summarise some of the insights from above: that the orthodox second law is violated on small scales was noted by Maxwell (1860). For him, even a weaker

[18]Some philosophers are more sceptical, such as Norton (2011).

[19]See (Ladyman et al., 2007) for a careful analysis and defence of Landauer's principle and (Maroney, 2009b) for a generalisation that includes non-deterministic logical operations.

version of the second law that prohibits the consistent and reliable transformation from heat into work could in principle be violated by a 'neat fingered being'. The insights from Smoluchowski and later Szilard, Bennett and Landauer were, however, that once the demon is modeled as a physical system, this weaker version of the second law is in fact not violated by demonic actions.

Maybe we should not be surprised that it took decades for physicists to get to the bottom of Maxwell's demon. With the demon, a new set of operations was introduced to thermodynamics, which only later was formulated in a mathematically and physically precise fashion: measurements, or in other words, the possibility of creating and exploiting correlations between a thermodynamic system and an external agent[20]. Traditional thermodynamics however is badly equipped to deal with such correlations; consider for example the following situation in which the position of a particle is to be measured by a memory cell, which is previously in it's ready-state (taken to be 'left') and which will align itself with the position of the particle after the measurement. Let us assume the particle actually is in the left side of the box. Schematically, the measurement process will then look like this:

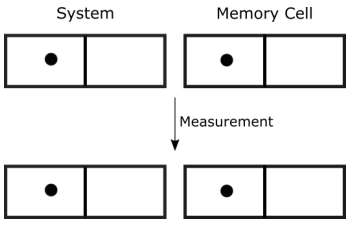

Figure 1.1: Measurement process of the position of an uncorrelated particle by a memory cell. The particle initially is in the left chamber and the ready-state of the memory cell is 'left' as well.

The physical situation is very different before and after the measurement step. From a thermodynamic point of view, however, nothing has changed. And yet we must

[20]Or automata. There is no need for the agent to be humanoid, it could be something as primitive as a memory cell.

acknowledge the fact that *something* has changed during the measurement, because the memory cell can reliably extract work from the system after, but not before, the measurement.

Now, if thermodynamics is about which processes an agent can or cannot perform on a thermodynamic system, then measurement outcome based operations (MOBOs) certainly ought to be included. However, the fact that thermodynamics itself treats systems as a black box requires further care in the case of MOBOs. In particular, the black box is now the joint system consisting of system and agent and what counts is the entropy balance of this joint system at the end of the thermodynamic cycle.

1.4 Classical Statistical Mechanics

This section will introduce various statistical mechanical notions of entropy for classical systems, without going into too much detail. The aim is to provide an overview over some of the conceptual difficulties that arise in the context of trying to identify the statistical notions of entropy with their phenomenological counterpart. Even though the main body of this thesis is concerned with black hole entropy, the von Neumann entropy and their relation to the phenomenological entropy, there are nevertheless several advantages to discussing entropy in a classical statistical mechanical setting before.

First, the reader will get an overview of the current philosophical debate and also an explanation of why 'knowledge' is often considered to play an important role. Understanding the origin of terms such as 'information' will help to evaluate their importance in a quantum and black hole setting respectively.

Second, the quantum statistical mechanical framework introduced in Chapter 3 is

based on the classical Gibbsian statistical mechanics. Understanding the conceptual and practical issues with the Gibbs entropy is hence useful for the task of determining whether these issues transfer to the quantum mechanical and black hole context. In particular it helps evaluating criticism against the quantum von Neumann entropy (Shenker, 1999; Hemmo and Shenker, 2006) and black hole thermodynamics (Dougherty and Callender, 2016; Wüthrich, 2017).

1.4.1 Entropies in Statistical Mechanics

This section will begin with some mathematical considerations about phase space, followed by the introduction of the Boltzmann and Gibbs entropy. For a thorough and historical overview over the developments of thermodynamics and statistical mechanics, I refer the reader to (Sklar, 1995).

We first consider a $6n$ dimensional phase space, where the instantaneous state of a system with $6n$ degrees of freedom is represented by a point in phase space with coordinates $(\mathbf{q}, \mathbf{p}) = (q_1, ..., q_{3n}, p_1, ..., p_{3n})$. This is called a *microstate*. The microstates form a complete and disjoint set on phase space. The evolution of the system through phase space is governed by the system's Hamiltonian. For convenience I will consider a *coarse-grained* phase space, in which each microstate is actually a small phase space volume element rather than a point. Strictly speaking, this is not true for classical systems, but everything below can be easily generalised to the fine-grained case.

We can also divide the state space system into non-degenerate, disjoint *macrostates*, each of which contains a set of microstates. The complete set of disjoint macrostates form a partition of the state space. Since the system is in only one microstate at any given moment in time, it is also in only one macrostate, for a specific partition, that

is for a particular way of dividing the state space into macrostates. However, it can of course simultaneously be in more than one macrostate if the macrostates belong to distinct partitions.

Let us now assign a *weight* ω_i to each microstate. The weight of the macrostate α is then the sum of the weights of the microstates it contains (Attard, 2002):

$$W_\alpha = \sum_{i \in \alpha} \omega_i. \tag{1.12}$$

These weights are non-negative real numbers, but not probabilities (they are not normed to one).

The weight of the total phase space is the sum of the weights of the macrostates (of a particular partition) or, equivalently, the sum of the weights of each microstate:

$$\Omega = \sum_\alpha W_\alpha = \sum_i \omega_i. \tag{1.13}$$

It is now possible to introduce the functional H. Acting on a particular macrostate, it is given by

$$H_\alpha = k_B \ln W_\alpha = k_B \ln \left(\sum_{i \in \alpha} \omega_i, \right) \tag{1.14}$$

where for now we leave the constant k_B unnamed[21]. The sum is taken over all microstates that lie within the macrostate α.

Acting on the entire phase space, that is on the sum of all macrostates, the functional takes the form

[21]Spoiler: k_B will of course the Boltzmann constant.

$$H = k_B \ln \Omega = k_B \ln \sum_i \omega_i = k_B \ln \sum_\alpha W_\alpha. \tag{1.15}$$

We can expand equation (1.15) by making use of the completeness of the macrostates $\sum_\alpha W_\alpha/\Omega = 1$ and by inserting an identity into the logarithm's argument:

$$H = \sum_\alpha \frac{W_\alpha}{\Omega} k_B \ln\left(\frac{W_\alpha \Omega}{W_\alpha}\right) \tag{1.16}$$

$$= \sum_\alpha \frac{W_\alpha}{\Omega} k_B \left[\ln W_\alpha - \ln\left(\frac{W_\alpha}{\Omega}\right)\right] \tag{1.17}$$

$$= \sum_\alpha \frac{W_\alpha}{\Omega} H_\alpha - k_B \sum_\alpha \frac{W_\alpha}{\Omega} \ln\left(\frac{W_\alpha}{\Omega}\right). \tag{1.18}$$

Depending on how we chose our partition, the weights of the individual microstates may all be equal, in which case we set $\omega_i = 1$, which means that then the weight of a macrostate simply consists of the number of microstates it contains, $W_\alpha = n_\alpha$. This however is a special case which disregards internal degrees of freedom a system may exhibit — typically these are ignored in classical phase space representations of gases.

Boltzmannian Statistical Mechanics We can now begin with an *interpretation* of equations (1.14) and (1.18). For this let us *assume* all microstates have equal weights, $\omega_i = 1$. A thermodynamic system, for example an ideal gas with n molecules, is then represented by a particular microstate, i.e. by a particular point in phase space. If we interpret equation (1.14) as the logarithm of the number of microstates, a particular macrostate contains, that is, if we associate a macrostate with a specific phase space *volume*, we arrive at the definition of the so-called *Boltzmann entropy*

$$H_B = k_B \ln W_\alpha, \tag{1.19}$$

The constant k_B then is the *Boltzmann constant*. A system is said to be at *equilibrium*, if its microstate lies in the equilibrium macrostate. This is usually the largest macrostate: "by far the largest volumes [of phase space] correspond to the equilibrium values of the macroscopic variables (and this is how 'equilibrium' should be defined)" ((Bricmont, 1995, p.179) cited in (Lavis, 2011, p.70)). The Boltzmann entropy is therefore maximal, if the system is at equilibrium. How to justify the association of the equilibrium macrostate with the largest macrostate is still a matter of debate, in particular since depending on the dynamics and constraints, the system may spend most of its time in non-equilibrium macrostates (Lavis, 2004, 2007). Current research by Werndl and Frigg (2015) suggests that one should associate the equilibrium state with the state the system *spends most time in* (Werndl and Frigg, 2015). But this is naturally a top-down argument: one already knows the properties one wants a system at equilibrium to have, in this case that it spends most its time there, and defines it accordingly. In particular, this presupposes the *approach* to equilibrium.

Since the time-reversible dynamics governing the movement of the system through phase space allows for the thermodynamic system to leave the equilibrium macrostate, i.e. to have a decrease in entropy, a justification for why a thermodynamic system ends up in the equilibrium macrostate needs to be given by Boltzmannians. Such a justification is usually given by arguments about *typicality*: a system out of equilibrium *typically* moves into the equilibrium macrostate. The notion of typicality is not unproblematic, in particular since it introduces probabilistic notions — which are exactly what many Boltzmannians do not want (why will become evident shortly). To sum up: the Boltzmannian regards entropy as a property of the macrostate the

system is in at a particular moment. This way, entropy is seemingly guaranteed to be an objective fact about a thermodynamic system, without any references to external agents.

Gibbsian Statistical Mechanics For the Gibbsian approach (Gibbs, 1878), we take the $\omega_\alpha/\Omega = p_\alpha$ to be *probabilities* over phase space. Equation (1.18) then reads

$$H = \sum_\alpha p_\alpha H_\alpha - k_B \sum_\alpha p_\alpha \ln p_\alpha. \tag{1.20}$$

If H_α is interpreted as the entropy associated with the individual macrostate α, the first term in (1.20) represents the average entropy per macrostate. The second term may be interpreted as providing a measure for the probability spread over the macrostates, which is maximal for equal probabilities and zero[22] for $p_\alpha \in \{0, 1\}$. So far, no assumptions about any particular partitioning into macrostates have been made, and Equation (1.20) is a general expression.

If we only take into account a system's *microstates*, equation (1.20) remains true. For equal weights[23] ($\omega_\alpha = 1$) the first term vanishes: $H_\alpha = k_B \ln \omega_\alpha = 0$. The entropy reduces to the well-known (discrete) *Gibbs* entropy

$$H_G = -k_B \sum_i p_i \ln p_i. \tag{1.21}$$

This expression for the discrete Gibbs entropy is formally equivalent to the information-theoretic *Shannon entropy* $H_{Sh}(X) = -\sum_n p(x_n) \ln (p(x_n))$ for a discrete probability

[22] We assume that $1 \cdot \ln(1) = 0$

[23] The underlying principle for assigning equal weights is called the principle of equal *a priori* probabilities. A mathematically and conceptually rigorous justification for this principle is difficult to obtain. Arguments usually rely on either ergodicity or time (volume, surface) averages, both of which justify equal probabilities for most, but not all cases.

source $p(X)$ with random variable $X \in \{x_n\}$. The Shannon entropy is often interpreted as a measure of negative information, or surprise[24]. As opposed to the Gibbs entropy, that reflects certain physical properties about a statistical mechanical system, the Shannon entropy is a property of a *probability source* and so the two entropies are conceptually distinct. This issue is distinct from the question of how we ought to interpret the probabilities.

In the Gibbsian approach, a system is at equilibrium, if the probability distribution is unchanging in time. This closely relates to the thermodynamic definition of equilibrium, where a system is considered at equilibrium if its thermodynamic variables do not change in time. But there is a big difference between a *probability distribution* not changing in time and actual thermodynamic variables not changing in time — in the Gibbsian setting, a system is still at equilibrium even if it fluctuates out of the equilibrium macrostate. For the Gibbsian, fluctuations therefore occur *at equilibrium*, whereas for the Boltzmannian they occur *in and out of equilibrium* (also pointed out in (Wallace, 2013a)).

I will now introduce two prominent criticisms against taking the Gibbs entropy as the correct statistical mechanical representation of the thermodynamic entropy. Both are often mentioned in the literature (Albert, 2003; Goldstein, 2001; Callender, 1999). The first one is of a technical nature and is based on the fact that if a system is at equilibrium (in the Gibbsian sense), then it must have always been there, due to the phase space volume conservation of the Liouvillian dynamics. This runs against our thermodynamic intuitions that systems *approach* equilibrium and remain there. This criticism is usually countered by introducing a *coarse grained* Gibbs entropy — however, how we justify this coarse graining is controversial, since justifications are

[24]But can also be given a physical interpretation as the average amount of bits needed to communicate a probabilistic event.

often given in epistemic terms[25], which leads us to the second criticism.

The second criticism regards the justification for the use of *probabilities*. In a highly influential account by Jaynes (1965), the probabilities occuring in the expression for the Gibbs entropy are taken to be measures of *ignorance*. This introduces an *epistemic* notion into the concept of entropy, running against the Boltzmannian (and thermodynamic) intuition that entropy is a property of the thermodynamic system in question and has nothing to do with lack of information[26]. Albert (2003) famously expressed his dismay with an epistemic approach to entropy:

> 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, 2003, p.64, original emphasis)

Understanding entropy in terms of an agent's external ignorance is certainly an intuitive option, given that the underlying dynamical equations of the system are deterministic. But there are other options. The next section will explore them.

1.4.2 Probabilities in Classical Statistical Mechanics

Proabilities are central to the Gibbsian account of statistical mechanics, and as pointed out above, their interpretation bears some conceptual difficulties. This section is aimed at giving a brief overview over the various ways to understand probabilities, especially in the context of statistical mechanics, before moving on to

[25] See Wallace (2013a) for more details on the different options.

[26] Note that it is equally possible to understand the Boltzmann entropy as a quantifier for an agent's uncertainty about the underlying microstate within a given macrostate.

the more interesting case of whether and how our notion of probabilities influences our conception of thermodynamic entities in statistical mechanics.

Making sense of the probabilities in classical statistical mechanics is highly non-trivial, as each of the various interpretations has its advantages and its draw backs. Interpretations of probabilities can be roughly categorised into 'ontic' and 'epistemic' interpretations. Ontic interpretations take probabilities to be "furniture of the world" (Frigg and Werndl, 2011), that is the probabilities exist independent of the perceiving mind of external observers. The two most prominent ontic interpretations are propensity interpretations and frequentism. In the first case, probabilities are taken to be intrinsic dispositions of a system (Popper, 1957, 1959), whereas in the latter case, probabilities are identified with relative frequencies of measurement outcomes. Propensity interpretations unfortunately hardly find attention in the philosophy of classical statistical mechanics literature, partly because there are some formal problems with propensities, such as Humphrey's problem (Humphreys, 1985), but also because propensities are taken to be hard to reconcile with the underlying deterministic dynamics of a classical statistical system. I will not go into more detail here, but some accounts trying to reconcile statistical mechanics and propensities, do exist (Clark, 2001). The other prominent ontic interpretation is frequentism. Frequentism was pioneered by Reichenbach (1971) and von von Mises (1936), and is the view that probabilities are to be identified with limiting relative frequencies of measurement outcomes. A problem is that even in the limit of infinite many measurements, the relative frequency may not be the actual probability, but instead fluctuate around it. Paradoxically, the limiting relative frequency will only give us the right probability with high probability. Frequentism by its nature is badly equipped to deal with probabilities of individual events. Not only do conceptual and practical issues arise (who has time to perform infinitely many measurements), but

also mathematical issues, as the brute identification of an actual infinite number of measurements with the probability requires the division of one infinite number by another (see for example (Wallace, 2012) for a discussion). These are general issues with frequentist accounts of probabilities, but there are also problems arising specifically in the context of statistical mechanics, if one is to identify the probabilities with long-term averages, i.e. identify phase-space average with time-averages. This requires the system to be *ergodic*, which is not guaranteed for all thermodynamic systems (Lavis, 2004, 2007).

Next to ontic interpretations, there also exist epistemic interpretations of probabilities. These identify probabilities with degrees of beliefs of an agent. Given the objective nature of scientific practice, will only be concerned with objective epistemic interpretations here, which means that two rational agents who share the same information ought to assign the same probability to an event Epistemic interpretations satisfy our intuition about probability–in particular in a setting in which the underlying dynamics is deterministic–of having something to do with our uncertainty about the future behaviour of the system, or alternatively our uncertainty about the exact setup of the experiment to be performed. The downside of an epistemic understanding of probabilities is that our credences may not necessarily reflect the actual physical situation, and so an agent may assign probability one half to a loaded coin, simply because she has incomplete information. This is a practical issue, but there exists also a conceptual issue which arises specifically in the context of statistical mechanics. If probability is a property of the observer's knowledge, then equilibrium also becomes a property thereof, at least in the Gibbsian setting. This runs against the intuition that there is something *physical* about a system being in equilibrium. In particular, in an epistemic setting, it seems as if the question of whether or not a system is at equilibrium can only be answered with respect

to the knowledge of an external observer. This equally applies to the approach to equilibrium (hence Albert's criticism above): the mixing of two gases who then settle into a joint equilibrium state is poorly explained by an agent's knowledge. I will come back to this issue shortly.

Even though probabilities are essential to the Gibbsian approach, the inability of getting a firm grip on the notion of probabilities is hardly something that is singular to statistical mechanics. Even in the seemingly harmless case of a coin toss there exists controversy about how we ought to interpret the probability assigned to the outcomes 'heads' and 'tails'. Just as in the case of statistical mechanics, the underlying dynamics of a (classical) coin toss is deterministic. As a result, both ontic and epistemic considerations appear to play a role in characterising the probability of landing heads or tails. There is an objective matter of fact about whether the coin is loaded or not, and any notion of probability should reflect this. But equally one may defend the epistemic dimension of the coin flip: whether or not the coin lands heads or tails is fully determined by its initial conditions and its dynamics and so any assignment of probability to the outcome must be based on an agent's uncertainty of these initial conditions or the dynamics. Probabilities occurring in a deterministic setting are always Janus faced[27]: they are neither purely epistemic, nor purely ontic, but somewhat both.

So far I haven't mentioned the notion of a statistical *ensemble*, a large imaginary collection of identically prepared systems, such that the probability of finding a system in a particular state agrees with the relative weight of all systems within the fictional ensemble in that particular state. The probability of finding the system in a certain state is then not directly associated with its relative frequency, but with its relative weight. As opposed to the relative frequency, the ensemble thereby offers an

[27]To my knowledge it was Prof. Harvey Brown who came up with this very fitting terminology.

opportunity to assign a probability to the single case. Probability, however, is not a property of a single system, but instead of a system and many other imaginary systems. Philosophers rightly tend to ask the question: "Why should the actual system care about its fictional twins?".

Ensemble interpretations are sometimes counted to the epistemic interpretations. However, they somehow defy classification as ontic or epistemic. Instead, they are probably better regarded as a tool that enables the physicist to talk about probabilities in the statistical mechanical context. Ensembles give an intuition about why measures on phase space should be conserved under Hamiltonian evolution, by taking into account all the "possible" micro-histories of a system in a given macrostate. But they hardly are satisfactory as explanation of what probabilities *are*. Nevertheless, in the context of statistical mechanics, the notion of a statistical ensemble shows up in almost any text book on the topic. The microcanonical and canonical probability distributions, referring to the probability distribution of a thermodynamic system at a fixed energy and a system in thermal equilibrium respectively, are better known as the microcanonical *ensemble* and the canonical *ensemble*.

It has been shown that the notion of probabilities in classical statistical mechanics is anything but unproblematic. Fortunately, the situation improves in the quantum context as will be shown in later chapters.

Inferential vs Dynamical Conceptions of Statistical Mechanics

A distinction between two different conceptions of classical statistical mechanics, which in particular is reflected by their different takes on probabilities, has recently been proposed by Wallace (2013a), who distinguishes between an *inferential* and a

dynamical conception. In the inferential conception, statistical mechanics is taken to be a quantification of an external agent's ability to make predictions about a system's behaviour, given a limited amount of knowledge of the system's state. The dynamical conception on the other hand refers to the idea that statistical mechanics involves giving an account of the actual dynamical behaviour of a system, unrelated to the presence of an agent or an observer. These two different attitudes, Wallace claims, are most naturally reflected in the conception of probabilities within statistical mechanics.

For an adherent of the dynamical conception, probabilities are not needed *a priori* - the initial conditions of a gas for example are fixed by certain macrofeatures, such as a certain fraction of molecules having a specific velocity (statistical considerations, according to Wallace). However, probabilistic considerations of some form are needed in order to make predictions about the macroscopic behaviour of the gas. In the case of the famous H-theorem for example it turns out that we can only predict the H-function to never increase *with high probability* but we cannot make statements about its definite dynamical behaviour (Tolman, 1938; Zermelo, 1896). The probabilities thereby are as previously stated not embedded into the description of the macroscopic features themselves but they do one way or the other occur in the description of the macro-*dynamics*.

For the *dynamicist*, the biggest challenge is to interpret these probabilities without slipping away into epistemic waters, after all, the system ought to be agent-independent. Probabilities therefore are to be understood as either long-term averages, relative frequencies in a fictional ensemble or as relative frequencies[28]. There are issues involved with each of these possibilities, as was shown above.

[28]Objective chances are one further option. However, given a deterministic microdynamics, objective chances are not the most natural choice.

The *inferentialist* will not need to commit to ontic interpretations of probabilities but may (and probably will) chose to understand probabilities as measures of epistemic uncertainty about the system's underlying microstate.

The supposedly largest challenge for her, according to some of the literature (Albert, 2003; Wallace, 2013a; Callender, 1999), is to make sense of the observed robustness and inevitability of the time-asymmetry in our world, namely that ice cubes melt and gas expands. Arguments usually go along the following lines: if probabilities simply refer to an agent's ability to predict the system's behaviour, since the Liouville measure is preserved under Hamiltonian evolution, then a system that moves from one equilibrium state into another (i.e. a gas expanding) will have the same entropy at its final stage than it had at its initial state. The Gibbs entropy (if seen as referring to epistemic probabilities) therefore cannot be interpreted as resembling the phenomenological thermodynamic entropy - which has increased. For the Gibbs entropy to be identifiable with the thermodynamic entropy the agent would need to discard *information*, contrary to the intuition that entropy reflects what is going on with the system.

Understanding entropy in a means-relative way, as discussed in the previous section, might potentially resolve some of this tension. If entropy quantifies how much work is extractable from the system given a set of fixed operations, then it becomes clear that the entropy must rise if the gas spreads out and the means remain the same. This then to a certain extent justifies the coarse-graining of the probability distribution when for example a partition is removed and the gas evolves into a larger phase space.

I now come back to the quote by Albert given at the end of the previous section, in he which expresses his dismay with the epistemic interpretation of the statistical

mechanical probabilities. The quote claims that an (epistemic) Gibbsian believes that her knowledge *brings about* the fact that milk dissolves in coffee. This in turn suggests a causal arrow from an agent's ignorance *to* the behaviour of the molecules, which is certainly nothing a serious physicist would ascribe to. Instead it is the other way around: *given* certain coarse grained initial conditions and *given* Hamiltonian dynamics, the system is expected to evolve into one of a set of possible states. The uncertainty associated with the question of which state the system *actually* evolves into is given by the probability distribution, but this has nothing to do with an ice cube not melting, if nobody is looking. Despite all this, the epistemicist will still have problems explaining the robust regularities that we observe in thermodynamic systems.

1.4.3 Three Reasons for using a Gibbsian Framework

For the remainder of the thesis, when I speak of statistical mechanics I will, unless specifically stated, use Gibbsian statistical mechanics as the correct statistical mechanical generalisation of the thermodynamic entropy. There are three reasons for this.

The first one is pragmatic: if we want to generalise entropy to quantum mechanics, a Boltzmannian approach seems unsuited for the task. In a quantum mechanical setting, a macrostate is uniquely described by the eigenvalues of mutually commuting observables $\hat{M}$ on the system's Hilbert space $\mathcal{H}_S$ (von Neumann, 1996). In a thermodynamic setting, these could be energy or particle number, for example. The Hilbert space can then be decomposed into subspaces $\mathcal{H}_{S,\alpha}$ in which the set of observables specifies a macrostate α. The quantum Boltzmann entropy $S_{q,B}$ is then given by (Greven et al., 2014):

$$S_{q,B} = k_B \dim \mathcal{H}_{S,\alpha}, \tag{1.22}$$

for a system in macrostate α. However, whereas in classical mechanics given a certain partition of phase space the system's microstate is found in uniquely one macrostate, this is not the case anymore in a quantum setting. The system may start out in a unique macrostate, but then evolve into a superposition of macroscopically distinguishable states, Schrödinger's cat states so to say. This then requires a revision of the concept of a macrostate and so the transition from classical to quantum is not as straight forward in the Boltzmannian case as it will be shown to be in the Gibbsian approach. In general, the Gibbsian approach is much more general than the Boltzmannian approach, allowing to statistically describe a much wider range of systems.

The second reason has to do with what I take to be an important task of philosophy of physics, namely to engage with the current scientific debate and identify conceptual ambiguities. In the context of quantum thermodynamics, the overwhelming majority of scientific work is done within a Gibbsian framework, but only few philosophers have thus far engaged with the conceptual foundations that underly the emerging fields of thermodynamic resource theories or single shot quantum thermodynamics, yet alone with the question of whether the difficulties one encounters in the classical context carry through to the quantum mechanical domain.

The third reason is that a lot of the criticism against the Gibbsian framework equally applies to the Boltzmannian approach, therefore giving the latter no clear advantage over the former. The Gibbs entropy is often criticised for being conceptually more problematic than the Boltzmann entropy and to furthermore not recover all of the desired properties of the thermodynamic entropy. As Callender (1999) showed,

however, *both* Boltzmann and Gibbs entropy are not able to fully recover all of the desirable attributes of the thermodynamic entropy and so in this regard, they are on equal footing. What about the conceptual feasibility of the Gibbs entropy? It was shown above that the question of how to best interpret the probabilities is difficult to answer. But the Boltzmannian approach faces very similar problems in this regard: in order to explain why entropy in ordinary thermodynamic systems is never observed to be decreasing, the Boltzmannian relies on the notion of typicality, i.e. that the Boltzmann entropy of a thermodynamic system is *typically* non-decreasing. To make this notion of typicality mathematically precise, however, the Boltzmannian will need to make use of probability measures herself[29]. And she will then face the same interpretational conundrum as the Gibbsian on how to interpret these probabilities. In this respect, therefore, the Boltzmannian approach does not fare much better than the Gibbsian approach. Furthermore, in the quantum case the problem of how to interpret probabilities becomes somewhat more tractable than in the classical case, as will be shown later on.

In light of the above three reasons, I will therefore continue with a Gibbsian understanding of statistical mechanics in mind.

[29]See also (Frigg, 2009) for a distinction between various notions of typicality.

Chapter 2

Black Holes

2.1 Introduction

> I believe that in order to gain a better understanding of the degrees of freedom responsible for black hole entropy, it will be necessary to achieve a deeper understanding of the notion of entropy itself. Even in flat spacetime, there is far from universal agreement as to the meaning of entropy — particularly in quantum theory — and as to the nature of the second law of thermodynamics. The situation in general relativity is considerably murkier [...]. (Wald, 2001, p.27)

In the last few decades, black holes have enjoyed an increasing amount of attention from physicists and philosophers alike, not least because of their exceptional status as objects for whose full description general relativistic, quantum theoretic and thermodynamic considerations seem to be needed. The surprising parallels between geometrical properties of black holes and classical thermodynamic variables have been taken to suggest that there exists a deeper connection between the laws of

black hole mechanics and the laws of thermodynamics. A connection that might go beyond mere analogy, possibly allowing us to *identify* the respective geometrical properties with their thermodynamic counterparts. This would then imply that black hole entropy, proportional to the black hole's event horizon area, is in fact identical to the thermodynamic entropy. With this identification remaining unchallenged by most physicists, a remarkable amount of effort is instead spent on finding a microphysical underpinning of the Bekenstein-Hawking (black hole) entropy, and with it an appropriate interpretation. However, until now no agreement on the matter is in sight (Wald, 1993; Susskind, 1993; Bombelli et al., 1986; Frolov and Novikov, 1998; Bekenstein, 2008) and so the pressing question remains: what is black hole entropy?

To make things worse, there is a remarkable amount of disagreement about the meaning of entropy even in cases where we *do* have a firm grip on the microphysics. Even if we deal with standard thermodynamic scenarios that do not involve black holes, it is far from clear which, if any, statistical entropy ought to be taken as the correct statistical mechanical generalisation of the phenomenological entropy, as was shown in the previous chapter. In addition to being conceptually distinct, the statistical mechanical candidates may even numerically disagree. In particular, none of them manages to recover all of the properties that are usually deemed necessary for the phenomenological entropy, as shown in (Callender, 1999) for the case of Gibbs and Boltzmann entropies. And so, even if one were to find an unproblematic microphysical derivation of the Bekenstein-Hawking expression, it would be far from clear whether this expression in fact would resemble thermodynamic entropy.

In the face of the amount of disagreement about the nature of entropy on the one hand, and the growing importance of black hole thermodynamics for the foundations

of physics on the other hand, I want to contribute to the debate about the nature of black hole entropy by showing that, within a given range of external parameters, black holes can in fact be considered genuine thermodynamic objects.[1] I will not do so by presenting yet another statistical mechanical argument for why entropy is proportional to horizon area, for the aforementioned reasons. Instead, I am interested in the question of whether we can consider the Bekenstein-Hawking entropy to be a *thermodynamic* entropy in the first place. I will therefore try to use only phenomenological reasoning, avoiding statistical arguments as far as possible. "As far as possible" because I take as given Hawking's 1975 result that black holes emit radiation at a temperature proportional to their surface gravity. This derivation of the Hawking radiation can of course not be done phenomenologically, but requires quantum field theory applied to curved space-time. It has hence ultimately a statistical (or rather quantum) origin. However, once established, the Hawking radiation behaves just like ordinary thermal radiation[2], whose behaviour uncontroversially can be described in a phenomenological thermodynamic setting.

The argument for a black hole entropy presented here differs from previous attempts, as it avoids controversial concepts such as 'information' and does not rely on the identification of a statistical mechanical entropy with the thermodynamic entropy. Instead, I will go back to the very beginnings of thermodynamics and use the notion of a Carnot cycle in order to derive the expression for the black hole entropy. The Carnot cycle is implemented by considering a box containing a black hole and a photon gas, playing the role of the working medium. By doing so, it will be shown that when a black hole is coupled to an uncontroversially thermodynamic object, i.e.

[1] A recent analysis by Wallace (2017) examines the same question from a slightly different angle but comes to the same conclusion.

[2] I am only concerned with an electromagnetic field in the vicinity of the black hole but my reasoning can be extended to imply other quantum fields as well.

the photon gas, the two of them combined behave thermodynamically and moreover entropy differences are proportional to differences of horizon area. This will provide strong reason to believe that black hole entropy is indeed thermodynamic entropy.

I will begin with a short recap of the laws of black hole mechanics before revisiting some of the arguments that have previously been made in an attempt to establish that black holes are thermodynamic objects and therefore have a thermodynamic entropy proportional to their respective horizon areas, as prominently made by Bekenstein (1973) and Hawking (1976). I will then present the black hole Carnot cycle and derive the expression for the thermodynamic entropy.

2.2 Preliminaries - Laws of Black Hole Mechanics

The most fundamental of law of thermodynamics, as was previously discussed, is the minus first law, which states that a system will spontaneously attain and remain in, unless subject to external change, a unique state of equilibrium, regardless of which state it was previously in (Brown and Uffink, 2001). In black hole mechanics, one may identify at least the latter part of this statement with the so-called *no hair theorem* (Misner et al., 1973). It states that a black hole at equilibrium, independent of the way it has formed, is uniquely characterised by a small number of parameters, namely its mass, charge and angular momentum.

The **zeroth law** of black hole mechanics states that the surface gravity κ of a static black hole is constant over the whole event horizon (Bardeen et al., 1973). The surface gravity of a black hole is given by the proper acceleration of a test particle near the event horizon multiplied by a redshift factor. That it is constant across the horizon is said to resemble the zeroth law of thermodynamics. However, it should

be noted that traditionally the zeroth law is concerned with the *transitive* relation between equilibrium states and it is merely a consequence of this relation that the temperature of a system at thermal equilibrium is constant throughout the system itself. The zeroth law of black hole mechanics therefore merely recovers a consequence of its thermodynamic counterpart, cheekily ignoring the far more challenging task of establishing transitive equilibrium.

The **first law** of black hole mechanics states that any change in black hole mass M needs to be balanced by a change in either surface area A and/or angular momentum J:

$$dM = \frac{1}{8\pi}\kappa dA + \Omega dJ, \tag{2.1}$$

where Ω is called the *angular velocity of the horizon* (Wald, 2001) and is constant. An extra term including an electrostatic potential and a change in charge can be added to the equation, but is of no importance for the argument, which is why it has been omitted. All of the above entities are defined for an observer at infinity, unless stated otherwise.

Equation (2.1) is taken to resemble the first law of thermodynamics,

$$dU = TdS - đW, \tag{2.2}$$

where dU resembles the change in energy, T the temperature, dS the change in entropy, and $đW$ the work done by the system. In thermodynamic textbooks, one often encounters $đQ$ for the heat term, instead of TdS. This is, because strictly speaking, T and S do not have their usual physical meaning at this point but are merely mathematical placeholders used to convert the inexact differential into an exact one. It is only by the second law that they obtain their canonical meaning as

absolute temperature and entropy, as discussed in the previous chapter.

The resemblance between equation (2.1) and equation (2.2) (the surface area sits in Equation (2.1) just as entropy sits in (2.2)), plus the fact that black hole horizon area originally was taken to be non-decreasing, led Bekenstein in 1973 to express his suspicion that the black hole horizon area could be interpreted as playing the role of the thermodynamic entropy[3]. He proposed the **generalised second law** of black hole mechanics to be the statement that the change of entropy in the black hole exterior plus the change of black hole surface area must be non-negative. At that point there still existed a problem about identifying the non-zero surface gravity of a black hole with what was thought to be a zero thermodynamic temperature. This issue was resolved when Hawking (1975) discovered that black holes are in fact not black but instead emit radiation at a temperature proportional to their surface gravity. It was Hawking's result which led to a precise definition of the black hole entropy as $S_{BH} = \frac{c^3 A}{4G\hbar}$. The generalised second law of black hole mechanics (Bekenstein, 1973; Hawking, 1976) then reads

$$\delta \left(S_{exterior} + \frac{c^3}{4G\hbar} A \right) \geq 0, \tag{2.3}$$

where $S_{exterior}$ refers to the entropy of the black hole exterior[4].

The third law of thermodynamics states that the entropy change of a system undergoing a reversible isothermal process approaches zero as the temperature at which that process is performed approaches zero kelvin. A consequence of this is that it is

[3]It is widely accepted that for a non-evaporating black hole a straightforward identification of the laws of black hole mechanics with the laws of thermodynamics fails as the temperature of such a black hole is necessarily zero at all times. See however Curiel (2014) for an opposing view on the matter.

[4]Curiously, Bekenstein (1973) calls this the 'common entropy' in the black hole exterior. It will be shown later, that 'common', for Bekenstein, does not mean 'thermodynamic'.

impossible to reduce the entropy of an object to zero within a finite number of steps. The **third law** of black hole mechanics is then the statement that it is impossible to achieve a surface gravity of zero within a finite number of steps. It will however play no role for the argument presented here.

2.3 Which Entropy?

2.3.1 Black Hole Entropy in the Literature

With the above in hand, it is undeniable that the laws of thermodynamics and the laws of black hole mechanics bear some extraordinary resemblance. However, in order to show that the laws of black hole mechanics simply *are* the laws of thermodynamics applied to black holes, more work needs to be done. In particular one must show that the entities referred to are in fact the same in both cases. This means demonstrating that some of the geometrical properties of black holes can indeed be identified with the thermodynamic variables. Here I will only consider the correspondence of the thermodynamic entropy with the black hole horizon area. As pointed out before, I will take it as uncontroversial that Hawking's 1975 result about particle creation at the black hole event horizon is correct, which allows us to identify the temperature of the electromagnetic quantum field outside the horizon with the ordinary black body temperature.

I will begin with a brief discussion of some of the already existing arguments that have been made in order to establish the analogy, if not identity, of entropy and horizon area. I then show how those results are unsatisfactory in their attempt to establish that black hole entropy is in fact genuine thermodynamic entropy. From herein I will refer to the identification of black hole entropy with thermodynamic

entropy as the BH-TD-identity.

One can divide the arguments in favour of an BH-TD-identity into broadly two types of approaches: on the one hand, there are those that appeal to the *similarities* in behaviour of horizon area and thermodynamic (sometimes statistical) entropy (SIM). The other group focusses on the *preservation of the second law of thermodynamics* (PRES). Arguments within PRES operate along the following lines: either one assigns a thermodynamic entropy to the black hole and associates it with the surface area, or the second law of thermodynamics can in principle be violated. Such approaches require the two entropies to be actually identical, as opposed to being mere analogues. SIM and PRES are related, of course, and neither of them would be convincing without the other: SIM can at most establish an analogy—but not an identity—between thermodynamic and black hole entropy. However, SIM does lay the conceptual groundwork for why we should suspect that there is any connection between the two entities at all. It is the task of the second approach to turn this suspicion into fact by establishing an actual physical link between black hole mechanics and thermodynamics.

In the following I will briefly consider the two first and arguably most influential authors to argue in favor of a BH-TD-identity.

Bekenstein

I will begin with Bekenstein, who in his influential 1973 article attempted a "unification of black hole physics and thermodynamics"(Bekenstein, 1973, p.2334). He begins his argument by pointing out some similarities between black hole area and thermodynamic entropy: the energy change of a black hole is as intimately related to a change of horizon area, as is (internal) energy change to a change of entropy in the

thermodynamic case. Furthermore, a merging of Schwarzschild black holes allows for the extraction of energy in the form of gravitational waves.Analogously, two thermodynamic systems, each individually at equilibrium, allow for work extraction when brought into thermal contact. These arguments all belong to SIM.

Bekenstein's further arguments are based on an information-theoretic interpretation of entropy, interpreted in terms of accessible information about the system's degrees of freedom.

> The connection between entropy and information is well known. The entropy of a system measures one's uncertainty or lack of information about the actual internal configuration of the system. (Bekenstein, 1973, p.2335)

Note that for Bekenstein, the information-theoretic entropy is more general than the thermodynamic entropy, even though the latter is to be understood in information-theoretic terms. A system's thermodynamic entropy, according to Bekenstein, measures the amount of uncertainty about the system's internal configuration, given few macroscopic parameters such as temperature, volume or pressure. In the case of black holes, this uncertainty is of a deeper kind, it quantifies the uncertainty about the black hole interior which in principle is inaccessible to an outside observer. Similar to the thermodynamic entropy, it equally is possible to characterise the state of the black hole by few parameters: the black hole mass, its charge and angular momentum (Misner et al., 1973). Furthermore, given that there are numerous ways the black hole could have formed, there must exist an inaccessible multiplicity of states for each set of parameters. This resembles the thermodynamic case, where each combination of macroscopic parameters can be realised by multiple configurations on the microscopic level. Entropy is then said to quantify this multiplicity. To give an (information theoretic) intuition of why entropy is related to surface area, Bekenstein

considers radiation or particles that fall into a black hole: information about the infalling objects is lost, but at the same time the surface area of the black hole increases, quantifying an increase in ignorance.

The above reasonings all appeal to SIM, the analogue behaviour of (thermodynamic and/or statistical) entropy and horizon area. But Bekenstein also uses PRES to argue for a black hole entropy.[5] If a violation of the second law is to be avoided, black holes need to have an entropy. Bekenstein illustrates this intuition by considering the following example. Stellar objects have a thermodynamic entropy. Once they reach the end of their lives, some of them collapse into black holes. If black holes didn't have an entropy, entropy effectively would have been destroyed during the formation of a black hole. This however constitutes a violation of the second law. It is this argument that lead Bekenstein to the fomulation of the generalised second law, which was already introduced in the previous section and which states that "The common [thermal] entropy in the black hole exterior plus the black hole entropy never decreases" (Bekenstein, 1973, p.2339).

In a recent paper, Wüthrich 2017 heavily criticises Bekenstein's argument:

> [...] the original argument by Bekenstein with its detour through information theory does not succeed in establishing the physical salience of the otherwise merely formal analogy between thermodynamic entropy and the black hole area, and so cannot offer the basis for accepting black hole thermodynamics as "the only really solid piece of information". (Wüthrich, 2017, p.16)

The reason for Wüthrich's criticism lies in what he describes as a flawed line of argument by Bekenstein:

> His [Bekenstein's] argument to this effect [...] is best analyzed as consisting of two parts. First, the thermodynamic entropy of isolated systems is identified

[5] It should be noted that there is some tacit assumption of information being conserved, whatever this means.

> with Claude Shannon's information-theoretic entropy. Second, the black hole area gets identified with the information-theoretic concept of entropy. Both steps combined then amount to an identification of the area of the black hole's event horizon with thermodynamic entropy. Since both steps do not just formally identify, but in fact carry physical salience across the identifications, the area of the horizon really is an entropy akin to the usual thermodynamic entropy. (Wüthrich, 2017, p.8)

Wüthrich's concern about Bekenstein's lighthearted identification of the information-theoretic entropy with the thermodynamic entropy seems *prima facie* justified[6]. But there is one caveat in Wüthrich's reasoning: Bekenstein never makes the claim that black hole entropy is thermodynamic entropy. In fact, he makes an effort to emphasise that black hole entropy is *not* thermodynamic entropy. In his 1973 article, Bekenstein writes: "[...] the black hole entropy we are speaking of is *not* the thermal entropy inside the black hole. In fact, our black hole entropy refers to the equivalence class of all black holes which have the same mass, charge and angular-momentum, not to one particular black hole." (Bekenstein, 1973, p.2336, original emphasis). This sounds more as if he has something like a statistical mechanical ensemble in mind. And indeed, seven years later Bekenstein again draws the distinction between thermodynamic and black hole entropy. This time he explains in more detail how he understands the notion of entropy.

> Thus, S_{bh} is something else than thermal entropy. In statistical physics, as well as in other fields, entropy has come to stand for missing information. With that interpretation, the thermal entropy of a cube of sugar simply measures our ignorance as to the precise microscopic states of the molecules in the cube, while its macroscopic state is fully described by chemical composition, temperature, volume and perhaps a few other variables. The ignorance of the

[6]Note that Wüthrich's criticism is just a specific example of the criticism already expressed above: *any* identification of the thermodynamic entropy with a statistical mechanical generalisation bears its problems. This does not only apply to the information-theoretic take on entropy but may equally reflect a general scepticism about about the reduction of thermodynamics to statistical mechanics.

> *exterior* observer about the matter inside a black hole is deeper. "Black holes have no hair" allows us only knowledge about M, Q and L for a stationary hole; neither the microstate nor composition [...], nor temperature, nor structure [...], nor anything else is–even in principle– measurable by a distant observer. The function S_{bh} is the obvious candidate for quantifying this deeper ignorance, so it is not surprising that it can be vastly larger than any reasonable estimate for the thermodynamic entropy. (Bekenstein, 1980, p.26)

In a plenary talk given at the seventh Grossmann meeting at Stanford, Bekenstein even calls entropy "one of the most abused terms in physics" (Bekenstein, 1994, p.3) and once more emphasises the relatedness, but not identity, of Boltzmann, Gibbs and Shannon entropy with the thermodynamic entropy.

From these quotes and from the remainder of his work, it becomes clear that when Bekenstein speaks about the black hole entropy, he is in fact *not* referring to the thermodynamic entropy. He does, in agreement with Wüthrich's claim, take thermodynamic entropy to be best understood in information-theoretic terms, but 'entropy' for Bekenstein is a much more general notion. And so equally the 'common entropy' outside of the black hole does not correspond to the 'thermodynamic entropy' outside, but in fact to the information-theoretic entropy of all possible degrees of freedom outside the black hole. Black hole entropy equally does not correspond to thermodynamic entropy, for Bekenstein, but still is a quantification of our ignorance, even if this ignorance is now of a different kind, due to the event horizon.

And so I conclude that Wüthrich's criticism against Bekenstein is largely unfounded: he does not rely on information-theoretic notions of entropy for his argument $S_{BH} = S_{TD}$, because he does not try to establish $S_{BH} = S_{TD}$ in the first place.

I will now move on to the second founding father of black hole thermodynamics, Stephen Hawking, and briefly introduce the argument he made in favour of a black hole entropy.

Hawking and thereafter

Stephen Hawking (1976) adopts some of Bekenstein's reasoning, but in addition provides his own argument for showing that the black hole entropy equals the statistical mechanical entropy. As opposed to Bekenstein, however, Hawking himself does not explicitly draw a distinction between the thermodynamic and the statistical entropy. He begins by assuming that there exists a finite amount σ of initial, uniformly distributed states that may give rise to a particular black hole. Just like in the ordinary statistical case, the entropy should be equivalent to $S_{BH} = \ln \sigma$. The entropy furthermore is required to be a function of M, Q and J. Hawking then demands that this entity always increase when matter or radiation falls into the black hole and that it is superadditive for two merging black holes. The only functions that satisfy the above criteria, as Hawking finds, are functions of the horizon area, the simplest of which is the area times a constant, which then turns out to be $c^3/4G\hbar$. If matter falls into the black hole, the change in values of M, Q and J leads to an increase in σ that's at least as large as the old value of σ times the number of configurations of the accretting matter. Hawking has therefore given a statistical mechanical derivation of what he takes to be the generalised second law.

Just like Bekenstein, Hawking takes $\ln \sigma$ to represent an agent's ignorance over the system's underlying micro-configuration. In his conclusion he writes: "The conclusions of this paper are that there is an intimate connection between [black] holes and thermodynamics which arises because information is lost down the hole." (Hawking, 1976, p.179).

Due to Hawking's lack of explicit distinction between the thermodynamic entropy and the statistical mechanical entropy, it remains a matter of speculation whether he

considers the two to be distinct[7] or not. Speaking in favor of him taking black holes to have a thermodynamic entropy is the fact that he in a later publication describes black holes as having "thermodynamic properties" (Hawking and Page, 1983, p.577), such as temperature and entropy. In this case, Hawking's argument presupposes the successful reduction of thermodynamics to statistical mechanics, which was shown to be problematic in the previous chapter of this thesis. As a derivation of S_{BH} being the *thermodynamic* entropy of black holes, Hawking's argument may be considered insufficient.

The distinction between thermodynamic and statistical mechanical entropy in the black hole literature has washed out significantly in the years following Hawking's article. In his book, Wald for example introduces the generalised second law as "the total entropy of matter outside black holes plus 1/4 the surface area of all black holes never decreases with time. This suggests that the laws of black hole mechanics literally *are* the ordinary laws of thermodynamics applied to a system containing a black hole" (Wald, 1992, p.55, original emphasis). For Wald, therefore, there is no doubt that the Bekenstein-Hawking entropy is just the ordinary thermodynamic entropy.

> [...] we must interpret S_{bh} as representing the *physical* entropy of a black hole, and that the laws of black hole mechanics must truly represent the ordinary laws of thermodynamics as applied to black holes. (Wald 2000, p.18, emphasis original)

He nevertheless also asserts that the question of how they arise from the underlying statistical mechanics is a mystery.

[7] In the sense of one being the generalisation of the other.

2.3.2 What the arguments don't show

I conclude so far that the arguments given by Bekenstein and Hawking do not establish that black holes are thermodynamic objects. The derivations of the generalised second law were always given by considering the statistical mechanical entropy. To obtain some clarity about the arguments given, it is nevertheless useful to discuss the two strategies SIM and PRES a bit further.

SIM, as pointed out earlier, appeals to the similar behaviour of black hole entropy and thermodynamic (or statistical) entropy. Such arguments can at most establish an analogy, but not an identity between the two entropies. Take as an illustration the case of 'similar' behaviour of a pendulum and an LC-circuit (made of a conductor L and a capacitor C): both are versions of a harmonic oscillator, but yet nobody would claim that spatial displacement is equal to electrical current. The majority of the arguments given by Bekenstein fall into this category and so they can at most establish an *analogy* between thermodynamics and black holes[8].

What about arguments that appeal to the preservation of the second law of thermodynamics? Aren't we forced to consider black holes having a thermodynamic entropy if we want to save the second law, one of the most well-established laws in physics? However, the above arguments either mix statistical and phenomenological entropies or merely appeal to the statistical entropy (Bekenstein, 1973; Hawking, 1976). But, as was already mentioned, the reduction of thermodynamics to statistical mechanics is not uncontroversial: it is far from clear *which*, if *any* statistical mechanical entropy is indeed the appropriate generalisation of the thermodynamic entropy. And so making use of some (shaky) underlying statistical mechanics to establish the generalised second law is problematic.

[8]To remain fair, this is all Bekenstein aims to achieve.

One may of course also question the whole enterprise: why should anyone care about whether black holes have a *thermodynamic* entropy that is proportional to their surface area? Statistical mechanics is more fundamental than thermodynamics, why do we not focus on the problem of deriving a statistical mechanical underpinning? As important as this last task is, there are nevertheless things to say in favor of establishing black holes having a thermodynamic entropy.

First, it is after all thermodynamics that gives physical meaning to temperature, heat and work and of course entropy. One of the tasks of statistical mechanics is to recover the laws of thermodynamics, but its range of application is naturally much broader than that. It is possible to assign a statistical mechanical entropy to the whereabouts of my bike keys, but this entropy is void of any (important) physical meaning. It tells me nothing about the thermal properties of my house. In particular, it does not allow me to talk about our familiar understanding of heat and temperature. Second, whereas in the classical case the microstates and the dynamics underlying the statistical mechanics is well known, the situation is much more complicated in the case of black holes. Here, a statistical mechanical description of the black hole is anything but business as usual and ought to be speculative in nature. It is little condolence that string theory has derived the Bekenstein-Hawking formula for black holes (Strominger and Vafa, 1996), given that it is unclear what physical meaning we ought to assign to it. In short: given that the statistical mechanical underpinning is unclear (see (Bekenstein, 2008) for a summary of all the attempts to derive the black hole entropy), it is a good idea to at least establish that black holes are thermodynamic in the first place.

Let us also briefly reconsider PRES as a strategy to show that black holes have entropy. In order to save the second law, the entropy of the black hole must rise by at

least the same amount as entropy of its surroundings was 'lost' (we may consider the example of a box of gas that falls into the black hole). But, if we do not require our system to return to its original state, then 'apparent violations' are nothing unusual. For example, it is not hard to come up with examples where heat is transferred from a colder to a warmer body, if system and environment are not required to return to their original state. Such examples would however not constitute a violation of the second law. The second law as presented in the previous chapter is best understood in terms of reversible *cycles*.[9] And so arguments that include thermodynamic objects to fall into the black hole are to a certain extend unsatisfactory, as they do not in an obvious manner allow us to construct a cycle.

2.4 Black Holes as Thermodynamical Objects

In the following I will present an argument which shows that black hole entropy can be considered to be actual, genuine, well-behaved thermodynamic entropy, given a range of external parameters. I will do so by considering a Carnot cycle with a black hole coupled to a photon gas as the working medium. In the next section, I will discuss what motivates such an approach.

2.4.1 Motivation

What is new?

The approach discussed here tries to avoid statistical notions, and in particular the concept of 'information', as much as possible. The goal is to show that black holes

[9]Some more recent approaches to phenomenological thermodynamics, in particular the axiomatic approaches of Lieb and Yngvason (1998) do not require the notion of cycles in order to derive the second law. I do not consider them here.

can indeed be considered to be true thermodynamics objects and to investigate whether their entropy is given by the Bekenstein-Hawking formula. To do so, the black hole will be coupled to an object which is uncontroversially thermodynamic in the sense that it behaves according to the laws of thermodynamics and that it can be described by the usual thermodynamic parameters. A photon gas at temperature T_g will take the role of this system. It will then be shown that the joint system can be used as the working substance of a Carnot heat engine. This approach differs from SIM, which considers the behaviour of isolated black holes. Instead, we say: if it really is a well-behaved thermodynamic system, then it must interact with other thermodynamic systems like a thermodynamic system does and both of them together must behave like a thermodynamic system[10].

There are a few assumptions necessary for the argument. One of them is that the second law of thermodynamics holds of our combined system. This is far from trivial and follows along the lines of PRES. I believe that it is nevertheless justified to do so, given that I show the existence of stable equilibrium states which allow us to consider reversible, quasi-static *cyclic processes*, such as first considered by Kelvin and Clausius, as will be discussed in the following.

Why cycles?

It was cyclic processes that inspired (or rather defined) the second law in the first place. Historically, the second law was formulated in terms of the efficiency of heat engines (Clausius): no cyclic process is more efficient than a reversible process. This emerges from the (phenomenological) fact that heat naturally always flows from warm to cold and never the other way round. From this it also follows that one cannot have a heat engine that (operating in a cycle) takes heat from a reservoir

[10]See Curiel (2014) for a justification of arguments of this sort.

and transforms it into work without producing any excess heat. As the efficiency of a reversible heat engine furthermore does not depend on the nature of the working fluid but is a function of the temperatures of the involved heat reservoirs only, one can define an absolute temperature scale.

Our main point of interest, thermodynamic entropy, enters the picture only now, as described in the previous chapter. From the Kelvin statement, one derives the so-called *Clausius inequality* $\oint đQ/T \leq 0$ with equality for reversible cycles, where $đQ$ is the heat flux into the system and T is the thermodynamic temperature of the heat reservoirs (the above equation is the limiting case of the discrete Clausius inequality $\sum_i đQ_i/T_i$, for which it is more obvious that the T_i indeed correspond to the temperatures of the involved heat reservoirs from which heat is extracted). The derivation of the Clausius inequality involves the cyclic process of a motor, which is instantiated by a series of Carnot heat engines that drive the motor, all operating between a principal heat reservoir and a number of auxiliary heat reservoirs. The motor will be back to its original state after a full cycle and together with the Kelvin statement of the second law, the Clausius inequality is derived. The temperatures occurring in the Clausius inequality all refer to the temperatures of the heat reservoirs. As $\int đQ/T$ is path independent, one can now define a state function, the entropy function, whose value (up to an arbitrary constant) solely depends on the state of the system. The Clausius inequality then becomes the entropy version of the second law $\Delta S \geq 0$ with equality for reversible processes in a thermally isolated system. The entropy can only be determined up to a constant.

In the approach given here, I will show that black holes behave like thermodynamic objects with a thermodynamic entropy given by the Bekenstein-Hawking formula. To do so, I will first show that a black hole coupled to a photon gas and enclosed in

a box behaves like a thermodynamic object insofar as that it can be used as a heat engine and perform a Carnot cycle. From the above discussion we can distill a few requirements that such a system necessarily must fulfill. First, each step must take place on a reversible path. Whether or not such reversible transitions are possible hinges crucially on the existence of equilibrium states: for a process to be reversible, the system needs to undergo quasi-static changes, at each instant of which the system is at equilibrium. A necessary requirement for a system to be able to undergo a reversible cycle is hence that it can be at thermal equilibrium for a range of external parameters. Furthermore, the existence of ideal heat baths is needed. It is with the temperature of the heat baths that we derive the absolute temperature scale.

It will shortly be shown that for the combined system, black hole, radiation and box, a thermodynamic temperature exists, but: does it follow that the black hole itself has a temperature? If we take Hawking's famous result for granted, then a black hole emits radiation with a thermal spectrum corresponding to some temperature T_{BH}. It is then possible for the black hole and the photon gas to be in a stable equilibrium state for which $T_{BH} = T_g$, as will be demonstrated. This potential for being in equilibrium with a thermodynamic object that *has* a temperature and being able to undergo reversible changes, allows us to say that the black hole equally *has* a temperature and not merely *emits* at a certain temperature. However, it should be mentioned that even though the black hole can be at equilibrium with the photon gas, the transitivity relation is restricted in the case of black holes. As Hawking (1976) remarked: a naked black hole could not be in stable equilibrium with an infinite heat bath due to its negative heat capacity. Even if the two started out in equilibrium, fluctuations would quickly lead to a run-away process, as will be explained in more detail shortly.

To summarize the above and before moving on to the meat of the argument: I will show that a black hole enclosed in a box with radiation gas can be used as a working substance for a Carnot cycle. Whereas we may expect a box with a piston and filled with gas to trivially be usable as a heat engine, it is not at all obvious whether we can expect the same when we include a strange stellar object such as a black hole. The main requirement for success is that the black hole can be in stable equilibrium with the gas and undergo a series of quasi-static transitions. It will be shown now, that this is indeed possible for a certain range of external parameters.

2.4.2 Equilibrium Conditions for Black Holes and Photon Gases

Black holes have a negative heat capacity of order $\propto -M_{BH}^2$, which contributes to a very counter-intuitive behaviour in thermal contexts. This means, that upon absorbing heat (or energy in general), black holes *cool down*. The first problem with having a negative heat capacity is that we cannot expect systems that start out at different temperatures, to ever be at equilibrium with each other. The second problem is that even if two systems start out at equilibrium, this equilibrium may not be robust against small fluctuations. As an illustration, consider a black hole initially in thermal equilibrium with a surrounding, infinite heat bath. Due to a fluctuation, its mass increases slightly, which in turn leads to a decrease in temperature. As its absorption cross section has now increased, it absorbs even more photons from the surrounding infinite heat bath. At the same time, as it has cooled down, its rate of emission decreases, and so the black hole will become even bigger and colder, and so on. The converse situation, in which the black hole fluctuates towards a higher temperature and a smaller mass, works analogously, leading to the complete

evaporation of the black hole. It was Hawking who first expressed this worry and arrived at the conclusion that "black holes cannot be in stable thermal equilibrium in the situations in which there is an indefinitely large amount of energy available." (Hawking, 1976, p.2).

Even though black holes cannot be in stable equilibrium with an infinite heat bath, they can be in stable equilibrium with a photon gas and enclosed in a box, for a certain range of parameters[11]. This then allows us to have the joint system undergo quasi-static transitions. Due to the negative heat capacity of the black hole, there nevertheless do remain some subtleties, which will discussed shortly. Placing the black hole in a box also allows us to extract work from the system in the old-fashioned way by having a piston which allows the volume of the box to vary or to be varied.

To begin, we take the box to be completely isolated with a constant total energy of

$$E_{tot} = E_g + E_{BH}, \tag{2.4}$$

where $E_g = \alpha V T_g^4$ is the energy of the photon gas within the box of volume V and at temperature T_g and $\alpha = (\pi k^2)^2/(15c^3\hbar^3)$. $E_{BH} = Mc^2$ the energy of the black hole with mass M.

We re-arrange equation (2.4) in such a way that the temperature of the photon gas is a function of the total energy, the black hole mass and the volume of the box:

$$T_g = \left[\frac{(E_{tot} - Mc^2)}{\alpha V}\right]^{1/4} \tag{2.5}$$

Hawking found that black holes radiate at the same rate as an ordinary body would if it were at a temperature inversely proportional to the black hole mass, namely at

[11] A similar discussion can be found by Custodio and Horvath (2003).

(Hawking, 1975)

$$T_{BH} = \frac{\hbar c^3}{8\pi k G}\frac{1}{M}. \tag{2.6}$$

At equilibrium, we require the black hole and the photon gas to be at the same temperature, $T_{eq} = T_{BH} = T_g$. Equating equation (2.5) and (2.6) and rearranging the terms leads to

$$f(M) = M^4(M - E_{tot}/c^2) + \beta V = 0, \tag{2.7}$$

where for convenience we introduce the constant $\beta = \frac{\hbar c^7}{15(8)^4\pi^2 G^4}$, (notably, this is not temperature). We can consider the above equation to be the *equation of state* for the system, black hole and photon gas. Equation (2.7) is equivalent to demanding that $\frac{\partial S_{tot}}{\partial M} = 0$ with $S_{tot} = S_{BH} + S_g$, but at this stage I want to avoid any entropy talk as much as possible.

Being a quintic equation, it's not always possible to analytically find the roots of $f(M)$. However, fortunately it is still possible to extract a sufficient amount of information from equation (2.7) that allows one to characterise the equilibrium conditions for the joint system.

Firstly, we can easily see that $f(M) \to \pm\infty$ as $M \to \pm\infty$. In addition, $f(M)$ has two turning points, namely at $M = 0$ and $M = \frac{4}{5}E_{tot}$, where we set $c = 1\frac{m}{s}$ for simplicity. Having two turning points means that $f(M)$ can have at most three roots. At zero mass, $f(M = 0) = \beta V > 0$, and since we know that for $M \to -\infty$, $f(M) \to -\infty$, one root must be at negative mass and hence irrelevant. Only the interval of $0 \leq M \leq E_{tot}$ is physical.

The existence of the two other equilibrium points depends on whether or not $f(M =$

$\frac{4}{5}E_{tot}$) is positive or negative (we require it to be negative), which can be rewritten as the following inequality[12]:

$$\frac{V}{E_{tot}^5} \leq \frac{0.082}{\beta}. \tag{2.8}$$

Equation (2.8) provides an upper bound for the volume of the box, given a fixed total energy (or alternatively a lower bound for the overall energy, given a fixed volume). This upper limit to the box size turns out to be sufficiently large: for a black hole of mass $M = M_\odot$ that accounts for 4/5 of the total energy the box may be as large as $1.5 \times 10^{30} m$.

If the volume V exceeds the limiting volume V_l, the only equilibrium that could be achieved is at $M = 0$, in which case the box contains only radiation but no black hole. Intuitively this makes sense: if we place a black hole into a box which is too large compared to the total energy, black hole and gas could therefore never equilibrate. If at some initial time $T_{BH} > T_g$, the black hole would absorb energy, grow and cool down, but the gas would also cool down and at a faster rate such that their temperature would never be equal. For $T_{BH} < T_g$ the opposite case would hold, leading to the complete evaporation of the black hole.

In Figure (2.1) we see two equilibrium points, one left (M_l) and right (M_r) of $M = \frac{4}{5}E_{tot}$. It is now essential to see whether those are stable and to show that a temperature difference between gas and black hole, $T_g < T_{BH}$ or $T_g > T_{BH}$ does not lead to a runaway process of the kind described above. One way to do so is by considering $f(M) = dS/dM$ and thereby interpreting the roots M_l and M_r as local extrema of the entropy function and fixing the energy and volume (Custodio and

[12]These values are consistent with those of Hawking (1976), who arrives at the same conclusion but with a statistical approach.

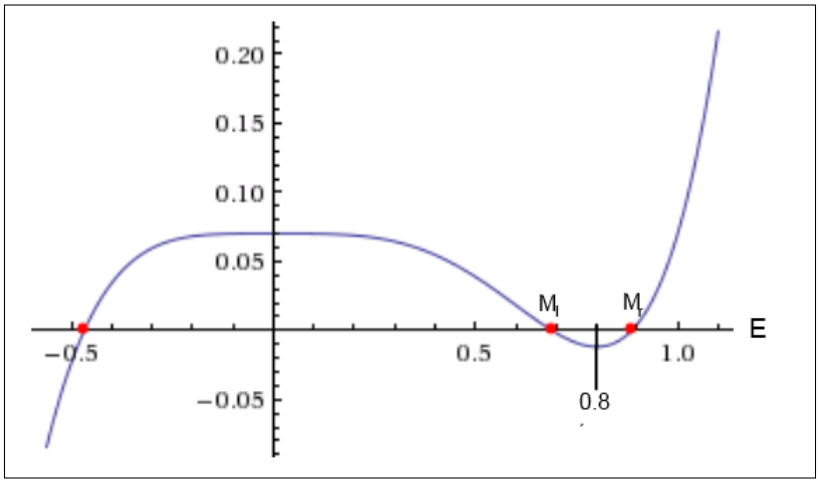

Figure 2.1: Exemplary graph for $f(M)$. The units on the x-axis are set to units of E_{tot}. The box volume used for this graph is $V = 0.07$, as can be seen from $f(M = 0) = 0.07$. We can see that two roots exist left (M_l) and right (M_l) of the turning point $M = 0.8E_{tot}$. For a volume larger than allowed by equation (2.8), these two roots don't exist. The root left on the far left is taken to be unphysical, as it refers to a negative black hole mass.

Horvath, 2003; Page, 1976) . It can then be shown that since $f'(M < \frac{4}{5}E) < 0$ and $f'(M > \frac{4}{5}E) > 0$, the root at M_l must be a stable local entropy maximum. The stability criterion can also be directly read off the graph in Figure (2.1).

This approach, however, presupposes the existence of an entropy function, which is what we want to avoid. It is still possible to show that the left equilibrium point, M_l, is stable by considering the rate of change in temperature upon a change in black hole mass. If it is the case that

$$\frac{dT_{BH}}{dM} < \frac{dT_g}{dM}, \tag{2.9}$$

then the gas can react quickly enough to any fluctuations in the black hole and there will be no run-away processes. Inserting all the relevant entities into the above equation, we get

$$-\frac{\hbar c^3}{8\pi G k}\frac{1}{M^2} < -\frac{1}{4\alpha V}\left[\frac{\alpha V}{E_{tot}/c^2 - M}\right]^{3/4}, \tag{2.10}$$

$$\frac{\left(E_{tot}/c^2 - M\right)^3}{M^8} > \frac{1}{4^4 \beta V}. \tag{2.11}$$

Applying the previous derived constraint for the box volume in the above inequality, it turns out that the condition is fulfilled for $M < 4/5 E_{tot}$ and so it is the left root M_l of $f(M)$ that denotes a stable equilibrium state.

We have therefore shown that (given a restriction on the ratio between total energy and box volume), a black hole can be in stable equilibrium with a surrounding photon gas[13], within the above defined parameter range. This means that for small perturbations, black hole and photon gas quickly equilibrate themselves again. For a Carnot cycle, transitions are taken to take place quasi-statically, and so at each time step the system has time to re-equilibrate itself.

With the above reassurance that a black hole can be in equilibrium with a photon gas when enclosed in a box of certain maximum volume, we now finally proceed with modeling a Carnot cycle. We will furthermore recover the well known expression for the black hole entropy from this analysis.

[13]Many thanks to David Wallace for suggesting a shorter derivation based on Thirring's stability condition (Thirring, 1970). For two systems A and B to be at thermal equilibrium, the following stability condition must be fulfilled: $\frac{c_A c_B}{c_A + c_B} > 0$, where c_i is the heat capacity of the respective system. If $c_A < 0$ and $c_B > 0$ then a necessary requirement for the two systems to be at thermal equilibrium is that $c_B < |c_A|$. This is another way to derive the upper limit to the box size.

2.4.3 Modelling a Black Hole Carnot Cycle

Setup

Just as above, we take as a working medium a black hole surrounded by a photon gas, enclosed in a box and attached to a piston. Both are at thermal equilibrium with each other at all times, namely $T = T_{BH} = T_g$.

The total energy of the system is given by

$$E_{tot} = U = Mc^2 + \alpha V T^4, \tag{2.12}$$

where M is the mass of the black hole, V the volume of the box and $\alpha = \frac{\pi^2 k^4}{15 c^3 \hbar^3}$ the usual radiation constant. One of the assumptions made earlier was that a photon gas can be considered a true thermodynamic object which is described by thermodynamic variables. It therefore has an entropy, which is

$$S_{rad} = \frac{4}{3} \alpha V T^3, \tag{2.13}$$

and exerts a pressure on the walls of the box of

$$P_{rad} = \frac{1}{3} \alpha T^4. \tag{2.14}$$

The box, containing both radiation and black hole, can be brought in contact with one of two heat baths at temperatures T_1 and T_2, where $T_1 < T_2$.

We now let the system go through the standard[14] Carnot cycle, which consists of

[14]Strictly speaking it is a reverse Carnot cycle. The standard Carnot cycle would be isothermal expansion, adiabatic expansion, isothermal compression, adiabatic compression. Since the system has negative heat capacity, however, this standard Carnot cycle would serve as a refrigerator. The

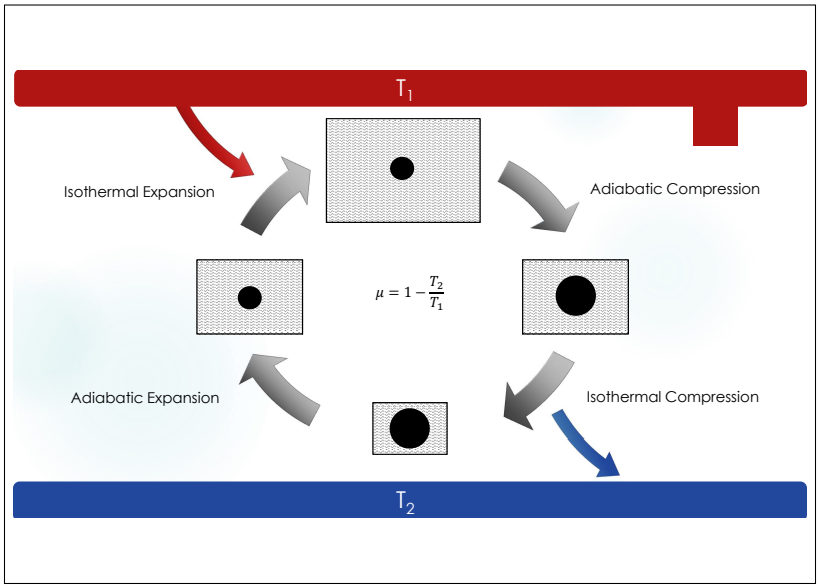

Figure 2.2: Schematic illustration of a black hole Carnot cycle. The system consists of a black hole and a photon gas, enclosed in a box. The size of the black hole is proportional to the temperature of the system, i.e. small is hot and large is cold.

isothermal expansion, adiabatic compression, isothermal compression and adiabatic expansion. An illustration of the process is given in Figure 2.2.

Isothermal Expansion

Since the overall system has a negative heat capacity, the notion of isothermal expansion during the Carnot cycle becomes a more subtle business than in the case of a system with positive heat capacity. In the latter case, a system is brought in contact with the hotter heat reservoir, and kept at the same temperature while expanding. Here, given its negative heat capacity, the system will be brought to a temperature just *below* the hot reservoir and kept there. When a small amount of heat flows from the reservoir to the system, the system will cool down very slightly. This is counterbalanced by simultaneously expanding the box, for which the system will

reverse Carnot cycle on the other hand acts as a heat engine, which is why we consider the reversed version.

need to do work, thereby heating up (due to its negative heat capacity). This way it is possible to expand the system from V_1 to V_2 and keep it at the same temperature, slightly below that of the heat bath. Since this temperature difference can be made arbitrary small, we still have effectively $T_{BH} \approx T_1$.

During this isothermal expansion the volume of the box changes from V_1 to V_2. The first law of thermodynamics

$$\Delta U_{12} = Q_{12} + W_{12} \tag{2.15}$$

holds (W_{12} is the work done on the system) and so we can calculate the heat flowing from the hot reservoir as

$$đQ_{12} = dU + PdV = \left(\frac{\partial U}{\partial T}\right)_V dT + \left(\frac{\partial U}{\partial V}\right)_T dV + PdV \tag{2.16}$$

The black hole mass only depends on the temperature of the black hole, and not on the volume of the box. Therefore, with the hot reservoir keeping the system at constant temperature $T = T_1$, we obtain for the total heat flux during the isothermal expansion of the box:

$$Q_{12} = \int_{V_1}^{V_2} \left(\alpha T_1^4 + \frac{1}{3}\alpha T_1^4\right) dV = \frac{4}{3}\alpha T_1^4 \left(V_2 - V_1\right). \tag{2.17}$$

The total system's entropy, which we have allowed to exist, since we must assume that the second law applies to the box system as a whole, must change during this process by

$$\Delta S_{12} = \int_1^2 \frac{đQ}{T} = \frac{4}{3}\alpha T_1^3 (V_2 - V_1). \tag{2.18}$$

Only the photon gas performs work and since its pressure only depends on the

temperature, the process is not only isothermal but also isobaric and we obtain for the work term

$$W_{12} = -\int_{V_1}^{V_2} P_{rad} dV = -\frac{1}{3}\alpha T_1^4 (V_2 - V_1). \tag{2.19}$$

The total change of energy of the system during this process is then

$$\Delta U_{12} = c^2 \Delta M_{12} + 4\alpha \left(V_2 T_1^4 - V_1 T_1^4 \right) = \alpha T_1^4 (V_2 - V_1), \tag{2.20}$$

Adiabatic Compression

During the adiabatic compression, the system does not exchange energy in the form of heat with its surroundings. The volume changes from $V_2 \to V_3$, but no heat is exchanged with the environment, $Q_{23} = 0$. The compression requires work, and so the total energy of the system changes and the system's temperature drops from T_1 to T_2. As the black hole mass is temperature dependent, its mass will now increase, $M_1 \to M_2$.

The adiabatic condition for a process with no heat exchange is given by:

$$\Delta U_{23} = W_{23}. \tag{2.21}$$

Using equations (2.12) and (2.21), we obtain the work done on the system during the adiabatic compression:

$$W_{23} = \gamma c^2 \left(\frac{1}{T_2} - \frac{1}{T_1} \right) + \alpha \left(V_3 T_2^4 - V_2 T_1^4 \right), \tag{2.22}$$

where we used the Hawking expression relating the temperature and mass of a black hole by $M(T) = \frac{\hbar c^3}{8\pi G k T_{BH}} = \frac{\gamma}{T_{BH}}$.

The first term of the above equation is ≥ 0 and the second term ≤ 0.

The work itself is only performed against the pressure of the photon gas, but the change of internal energy affects both the gas and the black hole at the center of the box. These must be equal, and so we differentiate equation (2.12) and equate it with the work done on the photon gas:

$$dU = c^2 dM + 4\alpha V T^3 dT + \alpha T^4 dV = -\frac{1}{3}\alpha T^4 dV. \tag{2.23}$$

Re-arranging the terms then leads to the following condition:

$$c^2 dM + 4\alpha V T^3 dT + \frac{4}{3}\alpha T^4 dV = 0 \tag{2.24}$$

The mass of the black hole only depends on the temperature, $M = M(T)$ and so $dM = \left(\frac{\partial M}{\partial T}\right)_V dT$. Equation (2.24) after some re-arranging of the terms then yields:

$$\frac{dV}{dT} + \frac{3V}{T} = -\frac{3c^2}{4\alpha}\frac{\partial M}{\partial T}\frac{1}{T^4}. \tag{2.25}$$

If we now insert the concrete expression for the black hole mass, $M(T) = \gamma/T$, equation (2.25) becomes

$$\frac{dV}{dT} + \frac{3V}{T} = \frac{3\gamma c^2}{4\alpha}\frac{1}{T^6}. \tag{2.26}$$

We solve this equation for the temperature-dependent volume, and so the above equation becomes

$$V(T) = -\frac{3\gamma c^2}{8\alpha T^5} + \frac{K}{T^3}, \tag{2.27}$$

where K is a constant. We can re-arrange the above in a useful way which resembles a bit more the traditional way of expressing adiabats:

$$VT^3 + \frac{3\gamma c^2}{8\alpha T^2} = \text{constant}. \tag{2.28}$$

Equation (2.28) describes the adiabats of the system. These are the paths of no heat exchange with the environment the state of the system follows during an adiabatic transition. Notably, they differ from the adiabats for a pure photon gas by the second term in equation (2.28).

Given that there is no heat exchange with the environment, the entropy change of the total system must be zero.

Isothermal Compression

The isothermal compression stage takes place analogously to the isothermal expansion stage, but this time the system is held at a temperature slightly *above* the temperature of the cold heat bath.

The amount of heat released is given by

$$Q_{34} = \frac{4}{3}\alpha T_2^4 (V_4 - V_3), \tag{2.29}$$

with an associated entropy change of the total system of $\Delta S_{34} = \frac{4}{3}\alpha T_2^3 (V_4 - V_3)$. The work performed on the system is

$$W_{34} = -\frac{1}{3}\alpha T_2^4 (V_4 - V_3), \tag{2.30}$$

and the total change in internal energy is

$$\Delta U = \alpha T_2^4 (V_4 - V_3). \tag{2.31}$$

Adiabatic Expansion

For the adiabatic expansion $V_4 \to V_1$, the adiabatic relations are given by

$$V_4 T_2^3 + \frac{3\gamma c^2}{8\alpha T_2^2} = V_1 T_1^3 + \frac{3\gamma c^2}{8\alpha T_1^2}. \tag{2.32}$$

Together with equation (2.28) we obtain the following relation:

$$V_3 T_2^3 - V_2 T_1^3 = V_4 T_2^3 - V_1 T_1^3. \tag{2.33}$$

The work done by the system during the adiabatic expansion is given by

$$-W_{41} = -\frac{\gamma c^2}{4}\left(T_1^2 - T_2^2\right) - \alpha K \left(T_1 - T_2\right). \tag{2.34}$$

Efficiency

The efficiency of a Carnot cycle is given by

$$\mu = \frac{W_{tot}}{Q_{12}} = \frac{W_{12} + W_{23} + W_{34} + W_{41}}{Q_{12}}. \tag{2.35}$$

The work terms of the adiabatic expansion and compression cancel each other out. We now derive the well-known Carnot efficiency relations by making use of the previously derived adiabatic relations $T_2^3 V_3 - T_1^3 V_2 = T_2^3 V_4 - T_1^3 V_1$ in the third step:

$$W_{tot} = Q_{12} + Q_{34} = \frac{4}{3}\alpha \left[T_1^4(V_2 - V_1) + T_2^4(V_4 - V3)\right] \tag{2.36}$$

$$= \frac{4}{3}\alpha \left[T_1 \left(T_1^3 V_2 - T_1^3 V_1\right) + T_2 \left(T_2^3 V_4 - T_2^3 V_3\right)\right] \tag{2.37}$$

$$= \frac{4}{3}\alpha \left[T_1 \left(T_1^3 V_2 - T_1^3 V_1\right) + T_2 \left(T_1^3 V_1 - T_1^3 V_2\right)\right] \tag{2.38}$$

$$= \frac{4}{3}\alpha \left[(T_1 - T_2) \left(T_1^3 V_2 - T_1^3 V_1\right)\right] \tag{2.39}$$

$$= \frac{4}{3}\alpha \left[T_1^4 (V_2 - V_1) \left(1 - \frac{T_2}{T_1}\right)\right]. \tag{2.40}$$

Together with $Q_{12} = \frac{4}{3}\alpha T_1^4(V_2 - V_1)$ we obtain an efficiency of

$$\mu = 1 - \frac{T_2}{T_1}. \tag{2.41}$$

This is the desired efficiency that we would expect from a Carnot engine. Our system thereby really can be used as a working substance for a heat engine.

2.4.4 Black Holes and Entropy

So far, it has been shown that black holes can be in equilibrium with a photon gas when enclosed in a box of certain maximal volume and furthermore, that the whole system can be used as working substance for a Carnot cycle by undergoing a series of quasistatic changes. In this section I will show how it is possible to calculate the black hole entropy just from these considerations (and without having had to make any assumptions about the thermal nature of a black hole beforehand). In the present case, a black hole was coupled to another thermodynamic system, a photon gas. To see why this is relevant, we consider again the entropy changes of the combined system. S_{tot} changed during the isothermal processes $1 \to 2$ and $3 \to 4$

but remained constant during the adiabatic processes $2 \to 3$ and $4 \to 1$. During the isothermal processes, the black hole did not change its state, and so the entropy change must have occurred exclusively in the photon gas.

During the adiabatic processes, however, S_{tot} remained constant, as no energy had been exchanged with the system environment. Nevertheless, between the black hole and the photon gas, energy *must have* been exchanged during the adiabatic transformations. This can be seen by comparing the adiabats of a photon gas in a box with those of our system comprising both photon gas and black hole: in the former case, the adiabats are given by $VT^3 = constant$. In the presence of a black hole, however, they are given by $VT^3 + 3\gamma c^2/4\alpha T^2$. This means that the thermodynamic path of no heat exchange of a solitary photon gas differs from that of our system. As a consequence, during the adiabatic compression, there *must* have been a heat flux out of the photon gas. And since the combined system is isolated and energy is conserved, the heat flux must be into the black hole. This sounds very much like Bekenstein's argument, but with one crucial difference: we have shown that black hole and photon gas can undergo a *reversible, quasi-static* cycle. This legitimises us in making use of the entropy formula $\Delta S = \int đQ/T$.

The entropy of the photon gas during such adiabatic compression therefore must *decrease*. Consequently the black hole needs to experience an *increase* of entropy. If the entropy is additive, and we believe that due to the reversible nature of the process this is a valid assumption, they cancel each other out exactly:

$$\Delta S_{Sys,23} = \Delta S_{rad,23} + \Delta S_{BH,23} = 0. \tag{2.42}$$

With the help of the adiabatic relations in equation (2.28), we can calculate the

entropy change in the photon gas to be

$$\Delta S_{rad,23} = \frac{4}{3}\alpha\left(V_3T_2^3 - V_2T_1^3\right) \tag{2.43}$$

$$= \frac{4}{3}\alpha\left(V_2T_1^3 + \frac{3\gamma c^2}{8\alpha T_1^2} - \frac{3\gamma c^2}{8\alpha T_2^2} - V_2T_1^3\right) \tag{2.44}$$

$$= \frac{\gamma c^2}{2}\left(\frac{1}{T_1^2} - \frac{1}{T_2^2}\right), \tag{2.45}$$

with the usual $\gamma = \frac{\hbar c^3}{8\pi G k}$.

For the above equation, we assumed that $\left(\frac{\partial M}{\partial T}\right)_V = -\frac{\gamma}{T^2}$ on the basis of Hawking's result for the black hole temperature. For an arbitrary $M(T)$, the change in photon gas entropy would instead read:

$$\Delta S_{rad,gen} = -\frac{\gamma c^2}{2}\int_{T_1}^{T_2}\frac{\left(\frac{\partial M}{\partial T}\right)_V}{T}dT. \tag{2.46}$$

If our black hole were not a black hole but instead a perfectly reflecting blob of mass $M(T) = M_{blob}$, the right hand side of equation (2.46) would vanish, resembling a zero entropy change of the photon gas. This is exactly what we would expect for the adiabatic transformation of a non-interacting photon gas.

Returning to our system in question, the entropy change of the photon gas as described in equation (2.45), must exactly counterbalance the entropy change in the black hole:

$$\Delta S_{BH,23} = \frac{\gamma c^2}{2}\left(\frac{1}{T_2^2} - \frac{1}{T_1^2}\right). \tag{2.47}$$

We can see that, up to a constant, the black hole entropy therefore must be of the form

$$S_{BH} = \frac{\gamma c^2}{2}\frac{1}{T^2} = \frac{\hbar c^5}{16\pi G k}\frac{1}{T^2}, \tag{2.48}$$

which is exactly the well-known expression for the black hole entropy. The horizon area scales with $\frac{1}{T^2}$ and so we have derived a black hole entropy that is proportional to the horizon area.

Given that we have coupled a black hole with an indisputably thermodynamic object and showed that the combined system behaves like a thermodynamic object and given that we have shown that there exist a range of accessible and stable equilibrium states between the subsystems, it seems reasonable to conclude that black holes really can be considered to be thermodynamic objects. For our analysis, we furthermore at no point needed to make reference to 'information' or statistical mechanical notions of entropy.

2.5 Conclusion and Outlook

I have shown that for a certain range of parameters, black holes can be taken to have a thermodynamic entropy given by the Bekenstein-Hawking entropy. However, the toy model of a black hole in a box is kept very simplistic and we ignored a large number of complicating factors that may play a role for a physically truthful analysis, including the effect of gravity on the photon gas or on the sides of the box, but we may expect the size of the box strongly to mitigate the significance of these effects. In particular, I only considered Schwarzschild black holes, which are crucial for establishing an equilibrium condition but whose existence in the universe is doubted by many. Still, the analysis presented here may be considered as a first step to a more rigorous treatment of black hole entropy. I have restricted myself to phenomenological thermodynamics without making use of statistical mechanical tools and thereby shown that, regardless of the relationship between statistical mechanics and thermodynamics, black holes have a thermodynamic entropy. Since

black holes sit at the interface of general relativity and quantum theory, taking a step towards a better understanding of black holes is also taking a step towards a better understanding of what is happening at this interface.

Finally, it should be noted that the idea of putting black holes into boxes filled with radiation has been proposed before, most prominently by Hawking (1976) and later by Custodio and Horvath (2003). Both established that stable equilibrium states between a black hole and a photon gas exist within a certain range of external parameters. However, in both approaches the argument is made by presupposing and appealing to the (statistical) entropy, which we will avoid here. A Carnot cycle cycle involving a black hole in a box has furthermore been suggested by Opatrný and Richterek (2011), where the authors use two black holes as heat sources/heat sinks respectively, therefore varying from this approach where the black hole is taken to be part of the working medium.

On the basis of the above results, there are a number of further steps for future investigation beyond the bounds of this thesis. In particular, a discussion of black hole entropy in the light of the adiabatic accessibility approach by Lieb and Yngvason (1998) would be fruitful in order to confirm our analysis. Furthermore, one may next look at similar arguments, such as the one brought forward by Curiel (2014), which claims that even classical black holes may be considered to be thermodynamic objects with a temperature. Another path worth taking is to look in more detail at the curious relationship between entropy and area. It turns out that this relationship is not unique to black holes but is indeed widely used in the context of spin chains and networks within the field of quantum information theory. Analysing this in more depth may result in a better understanding of black hole entropy.

Chapter 3

Quantum Mechanics

In this chapter I will explore the notion of entropy in the quantum context. The chapter will be divided into three parts, each of which explores a different facet of entropy and thermodynamics when applied to quantum system. The three parts consist of a discussion of von Neumann's original argument for a quantum entropy, a broader discussion about the relation between thermodynamics and quantum mechanics and an excursion into a recent claim made by Cabello et al. that thermodynamics allows us to make predictions about the tenability of some interpretations of quantum mechancis.

In the first part, I will present von Neumann's original argument in favor of an identification of $-\mathrm{Tr}\rho\log\rho$ (nowadays called the von Neumann entropy) with the thermodynamic entropy (von Neumann, 1996). The validity of his argument has been contested by Shenker (1999) and Hemmo and Shenker (2006), who claim to have found an undermining counterexample to von Neumann's argument. I will show that their counterexample is itself technically and conceptually flawed to the extend that it cannot provide us with any new insights about the relation between

the thermodynamic entropy and the von Neumann entropy.

The second part attempts to reconcile thermodynamics and quantum statistical mechanics by giving a detailed account of how the thermodynamic laws can be derived from the fundamental principles of quantum mechanics and a set of explicitly stated assumptions. This derivation follows the work of Maroney (2007) and makes use of a range of well known results in statistical mechanics[1]. In this approach, the identification of the canonical ensemble with the thermal states encountered in thermodynamics is based on phenomenological considerations about so-called *passive* states (Pusz and Woronowicz, 1978; Lenard, 1978; Sewell, 1980) and the existence of heat reservoirs. This second part will also touch base with contemporary research in quantum thermodynamics. The newly emerging fields of quantum resource theories (Brandão et al., 2013; Gour et al., 2015) and single shot quantum thermodynamics (Rio et al., 2011; Horodecki and Oppenheim, 2013; Skrzypczyk et al., 2014; Brandão et al., 2015) will be contrasted with the present approach and with phenomenological thermodynamics itself, providing insight into the potentials these fields bear for the foundations of thermodynamics.

In the last part of this chapter I will examine an exciting attempt by Cabello et al. (2016) to use thermodynamic reasoning and apply it to the very foundations of quantum mechanics: the question of how to interpret quantum mechanics. The authors claim that there is an empirically verifiable difference between two sets of quantum interpretations and that one of them ought to be considered untenable. The argument is based on results from computational mechanics. It will be shown that their conclusion does not follow and that their surprising result can be traced back to the illicit omittance of the entropy contribution of an external agent.

[1]To be found for example in (Gibbs, 1878; Szilard, 1925; Tolman, 1938; Wehrl, 1978; Partovi, 1989).

3.1 Von Neumann's Argument for a Quantum Entropy

This chapter will now begin by discussing von Neumann's argument for $-\mathrm{Tr}\rho\log\rho$ being the correct quantum mechanical generalisation of the thermodynamic entropy and the criticism expressed by Shenker (1999) and Hemmo and Shenker (2006).

3.1.1 The Argument

In his seminal book, von Neumann (1996) presents an argument in which he determines the entropy of a quantum mechanical ensemble with density matrix ρ and establishes that entropy is non-decreasing in both type 1 and type 2 processes[2].[3] In his argument, he considers the cyclic transformation of a quantum gas confined to a box. By demanding that the overall entropy change of system and heat bath must be zero by the end of the cycle, von Neumann concludes that this is only possible if the entropy of the quantum gas is given by $-\mathrm{Tr}\rho\ln\rho$.

Von Neumann's argument is based on a Gedankenexperiment that considers a quantum gas consisting of systems $[\mathbf{S}_1, ..., \mathbf{S}_N]$ where each quantum system is locked up in a box $\mathbf{K}_1, ..., \mathbf{K}_N$. Each box is heavy and shields off the contained quantum system from the environment, blocking any possible interactions with the other systems. The boxes are then placed in an even larger box $\bar{\mathbf{K}}$ of volume V, which is significantly larger than the small boxes and which is in thermal contact with a heat bath. Von Neumann then assumes that this quantum gas initially is in a

[2]Von Neumann considers two types of processes that describe how the quantum state changes in time. The first, 'Prozess 1', is associated with the probabilistic outcome of a measurement, whereas 'Prozess 2' refers to the evolution of the system via the Schrödinger equation.

[3]He explicitly assumes the validity of the second law for this.

mixed state[4] $\rho = \lambda\,|+\rangle\langle+| + (1-\lambda)\,|-\rangle\langle-|$, where the vectors $|+\rangle$, $|-\rangle$ are required to be orthonormal[5] I will restrict this discussion to a mixture of $\text{rank}(\rho) = 2$. A generalisation to n dimensions is straight forward. . Orthogonality is a requirement for the existence of semi-permeable membranes that play an important role in von Neumann's thought experiment. Such membranes are permeable to one state but impermeable to another. To complete the setup, another box $\bar{\mathbf{K}}'$ of equal volume V but empty inside, is added to the left of $\bar{\mathbf{K}}$. The walls between $\bar{\mathbf{K}}$ and $\bar{\mathbf{K}}'$ are then exchanged for a semi-permeable membrane of the above kind. It reflects systems in state $|-\rangle$ but allows systems in state $|+\rangle$ to pass. From the right hand side of box $\bar{\mathbf{K}}$ another semi-permeable membrane is now pushed in, which is permeable for $|-\rangle$-systems but impermeable for $|+\rangle$-systems. This way the $|+\rangle$-systems are 'pushed' into the left box, $\bar{\mathbf{K}}'$. At the end of this process, the previous mixture will be separated into a $(+)$- and a $(-)$- gas and the separation will have taken place without performing any work and without any heat exchange with the heat bath.

In the next step, the two boxes are compressed to volumes λV and $(\lambda-1)V$ respectively, while keeping the temperature T constant, which changes the densities in the boxes from $\lambda N/V$ and $(\lambda-1)N/V$ to the initial density of the gas of N/V, where N is the number of systems, or molecules. The $|+\rangle$- and $|-\rangle$-gases are then reversibly transformed into a $|\psi\rangle$-gas via unitary operations and the partition is removed. During the isothermal compression, the entropy of the heat reservoir increases by $Nk_B\lambda\ln\lambda$ and $Nk_b(1-\lambda)\ln(1-\lambda)$ respectively. Von Neumann argues that since the whole process is reversible, the total entropy change of gas and reservoir must be 0 and since the (normed) entropy of the final $|\psi\rangle$-gas is 0 by definition, it must have been $S = S_+ + S_- = -Nk_B\left[\lambda\ln\lambda + (1-\lambda)\ln(1-\lambda)\right]$ before. In a later

[4]Notably, von Neumann does not give specifications about the size of the quantum systems inside the boxes $\mathbf{K}_i$. However, he regards the boxes themselves as acting like 'molecules'.

[5]Mixtures instead of pure states are also conceivable, as long as they are disjoint.

discussion he also explains that his type 1 process leads to increase in entropy.

A schematic illustration can be found in Figure 3.1.

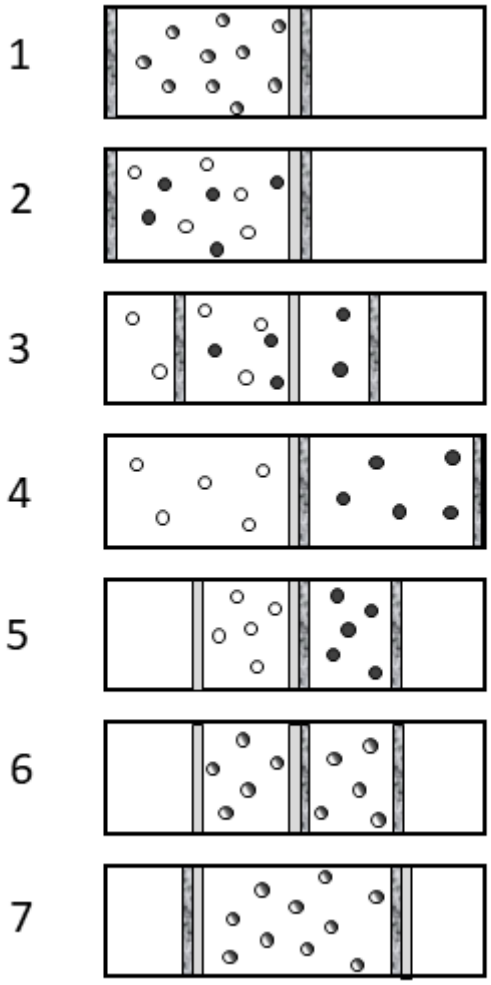

Figure 3.1: An illustration of the argument. Von Neumann's argument itself begins at Stage 2. **(1)** The individual particles are each in a superposition $|+\rangle$ and $|-\rangle$, indicated by bicoloured circles. **(2)** After the spin measurement, the system is now in a mixed state. The particles are either in spin-up (white) or spin-down (black) states. **(3)** The spin-up particles are separated from the spin-down particles with the help of semi-permeable membranes. **(4)** The spin-up particles are now completely separated from the spin-down particles. **(5)** The two chambers are compressed to half of their respective volumes. **(6)** The $|+\rangle$- and $|-\rangle$-gases are now transformed back into their original state of superposition. **(7)** The partition is removed. In this figure, no heat bath is present, but we can imagine that it exists and takes up the dissipated entropy at $4 \rightarrow 5$.

We can generalise the above considerations to more dimensions and consider a density matrix ρ with eigenvectors $|\phi_1\rangle, ..., |\phi_n\rangle$ and eigenvalues $\lambda_1, ..., \lambda_n$. The entropy is then given by $S_\rho = -\text{Tr}\rho \ln \rho = -\sum_{i=1}^{n} \lambda_i \ln \lambda_i$.

3.1.2 Shenker's (1999) Criticism and Henderson's (2003) Reply

I will now briefly consider Shenker's (1999) first criticism against von Neumann's argument and Henderson's (2003) reply. According to Shenker, two assumptions were made by von Neumann:

a) the thermodynamic entropy only changes at the compression stage $4 \to 5$, and

b) the entropies of stage 1 and 7 are the same.

As Shenker presents the argument, von Neumann's conclusion was that the entropy must have increased during the measurement process at stage $1 \to 2$, to balance out the decrease at step 4. The discrepancy between the behaviour of the von Neumann entropy and the "classical entropy"[6] allegedly becomes apparent when considering stages 2 to 4. We first consider the change in von Neumann entropy: at stage 2, the system is in a maximally mixed state and therefore has a positive von Neumann entropy. At stage 4, Shenker claims, the system is in a pure state and therefore has zero von Neumann entropy by definition. The von Neumann entropy hence must have decreased between 2 and 4.

[6]Shenker does not (yet) distinguish between classical statistical mechanical entropy and thermodynamic entropy: "Classical thermodynamics concludes that the very separation [of two gases] means a reduction of entropy. This is called entropy of mixing [...]" (Shenker, 1999, 39). Entropy of mixing however has its origin in statistical and not phenomenological considerations.

"From a thermodynamic point of view" (p.42), however, the entropy has not changed from $2 \rightarrow 4$. This is because "the entropy reduction of the separation is exactly compensated by an entropy increase due to expansion" (Shenker, 1999, 45). The thermodynamic entropy, according to Shenker, only changes between $4 \rightarrow 5$. Thermodynamic entropy and von Neumann entropy therefore differ in their behaviour since a reduction in thermodynamic entropy takes place at stage 5, as opposed to the reduction of von Neumann entropy at stage 4.

Henderson (2003) rightly points out some deficiencies in Shenker's argument that explain the alleged discrepancy. She shows that the system at stage 4 *cannot* be considered to be in a pure state, the gas' spatial degrees of freedom must also be taken into account in addition to its spin degrees of freedom. The initial state of the system at stage 1 is then given by $|\Phi\rangle \otimes \rho_\beta$, where ρ_β is the thermal state of the system in contact with a heat bath at inverse temperature β. The entropy change at stage 2 is then only due to the entropy change of the spin degrees of freedom. Furthermore, even if we assume collapse, as Shenker implicitly does, the entropy is still high, since "we lack knowledge of *which* pure state the system is in" (Henderson, 2003, 294, original emphasis). The separation step at stage $2 \rightarrow 4$ then only 'labels' the states in so far as they are associated with a particular spatial area of the box, but this step does *not* change the entropy. The change in entropy at the compression stage 5 is then due to a change of the entropy of the *spatial* degrees of freedom.

3.1.3 Modern Criticism by Hemmo and Shenker (2006)

In a subsequently published, revised and amended version, Hemmo and Shenker (2006) offer an amended proposal with a similar but slightly weakened claim. They assert that "von Neumann's argument does not establish a conceptual link between

$-\mathrm{Tr}[\rho \ln \rho]$ and the thermodynamic quantity $(1/T) \int pdV$ (or dQ/T) *in the single particle gas* [...]" (Hemmo and Shenker, 2006, p.158, emphasis added). They therefore retain their position that the von Neumann entropy cannot be empirically equivalent to the phenomenological entropy, but restrict this inequivalence to the domain of single or sufficiently few particles. Von Neumann and thermodynamic entropy, they argue, are effectively equivalent only in the thermodynamic limit.

This section will discuss Shenker's and Hemmo and Shenker's (H&S) efforts to show dissimilar behaviour between the two entropies and reveal that their argument is flawed to the extent that it allows for *pepetua mobile* of the second kind. The source of error will be identified to be the failure to take into account the entropy contribution of the measurement apparatus. It was already demonstrated by Szilard (1929) and his famous one-particle engine, that measurement based correlations with an external agent cannot be ignored in the single particle limit, as they straightforwardly lead to a violation of the second law. Once the entropy contribution of the measurement apparatus is taken into account, however, the analogous behaviour of thermodynamic entropy and von Neumann entropy for the *joint system* is restored.

I assume, just like H&S, that it is in fact possible to treat a single quantum particle as a genuine thermodynamic system (see Chapter 1 for a justification).

The following argument, including any conceptual ambiguities, has been taken unamended from (Hemmo and Shenker, 2006). An illustration of the steps can be found in Figure 3.2.

Step 1 (Preparation I): A quantum particle is prepared in a spin-up eigenstate in the x-direction, $|+_x\rangle_P$. Its coarse-grained initial location is given by $\rho(L)_P$, where the subscripts L and R will refer to its position in either the left or the right part of the box. The measuring apparatus M starts out in the state $|\mathrm{Ready}\rangle_M$. The initial

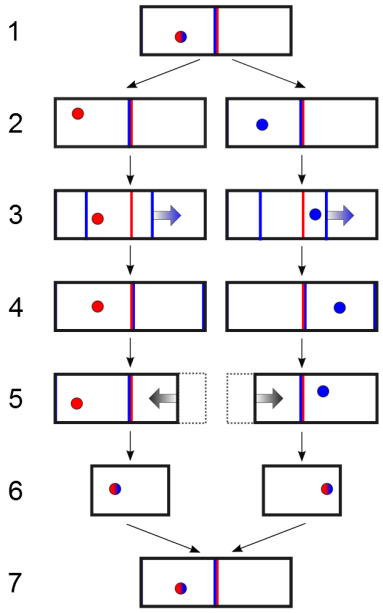

Figure 3.2: Illustration of the Gedankenexperiment following Hemmo and Shenker (2006). (1) the particle is prepared in a spin-x up eigenstate. (2) A spin-z measurement is performed on the particle. (3) Depending on the outcome, the particle is moved to the right side of the box or remains in the left side via semi-permeable membranes. (4) A location measurement is performed. (5) The empty side of the box is compressed. (6) & (7) The system is brought back to its original state.

state of the particle is hence given by the product state:

$$\rho^{(1)} = |+_x\rangle \langle +_x|_P \, \rho(L)_P \, |\text{Ready}\rangle \langle \text{Ready}|_M \, . \tag{3.1}$$

Step 2 (Preparation II): In Step 2, a measurement in the spin z direction is performed, leading to an entanglement of the measurement apparatus' pointer states and the z spin eigenstates. It is important to note that H&S do not specify the nature of the measurement at this stage, i.e. whether they are working in a collapse or no-collapse model. The state of the system is given by

$$\rho^{(2)} = \frac{1}{2} \left(|+_z\rangle\langle +_z|_P \, |+\rangle\langle +|_M + |-_z\rangle\langle -_z|_P \, |-\rangle\langle -|_M \right) \rho(L). \tag{3.2}$$

The reduced density matrix of the quantum system becomes:

$$\rho^{(2,red)} = \frac{1}{2}\left(|+_z\rangle\langle+_z|_P + |-_z\rangle\langle-_z|_P\right)\rho(L), \tag{3.3}$$

which, as H&S state, "in some interpretations may be taken to describe our ignorance of the z spin of P" (Hemmo and Shenker, 2006, p.160).

Whereas the von Neumann entropy of the spin component $S_{vN} = -\mathrm{Tr}[\rho \ln \rho]$ was zero before, it now becomes positive. The thermodynamic entropy however, the authors assert, remains the same.

Step 3 (Separation): Two semi-permeable membranes are inserted and moved through the box in a way that the particle remains on the left iff it is in state $|+_z\rangle$ but is moved to the right iff it is in state $|-_z\rangle$. There is no work cost involved in this process and neither von Neumann entropy nor thermodynamic entropy change during this step, during which the spatial degrees of freedom are coupled to the spin degrees of freedom.

Step 4 (Measurement): As we are only considering a *single* molecule in this setup as opposed to von Neumann's original many particle gas, the compression stage needs to be preceded by a location measurement in order to determine which part of the box is empty. H&S therefore introduce a *further* measurement before compression (not present in von Neumann's original argument), in order to determine in which part of the box the particle is located.

H&S add that for the calculation of the von Neumann entropy, collapse and no-collapse interpretations will now have to use different expressions for the quantum state. In collapse theories, the state as a result of the location measurement collapses into either

$$\rho^{(4,+)} = |+_z\rangle \langle +_z|_P \, \rho(L) \quad \text{or} \quad \rho^{(4,-)} = |-_z\rangle \langle -_z|_P \, \rho(R). \tag{3.4}$$

For no-collapse interpretations on the other hand, the system's state is given by the reduced density matrix:

$$\rho^{(4,red)} = \frac{1}{2} |+_z\rangle \langle +_z|_P \, \rho(L) + \frac{1}{2} |-_z\rangle \langle -_z|_P \, \rho(R). \tag{3.5}$$

The thermodynamic entropy, S_{TD}, as the authors stress, is *not* influenced by the position measurement and does not change during this step, in the sense–presumably– that no heat flows into, or out, of the system in consequence of this measurement . By contrast they urge, whether the von Neumann entropy changes, depends on whether we consider collapse or no-collapse interpretations. In the case of collapse interpretations the von Neumann entropy of the system allegedly decreases, whereas in the case of no-collapse interpretations, it remains the same.

Step 5 (Compression): An isothermal compression of the box back to its original volume V is performed. The change in *thermodynamic* entropy during this step is normally given by $S_{TD} = (1/T) \int pdV$, however, since there is no work involved in the compression against the vacuum, H&S argue, the thermodynamic entropy does not change at Step 5. In fact, the thermodynamic entropy does not change throughout the *whole experiment*, the authors claim.

Step 6 (Return to Initial State): The system is brought back to its initial state by unitary transformations with no entropy cost. "[...] [T]he measuring device need also be returned to its initial ready state. One can do that unitarily." (Hemmo and Shenker, 2006, 161).

H&S's main criticism thereby focuses on the fact that the thermodynamic entropy

remains constant *throughout the experiment*, whereas the von Neumann entropy does not:

> Therefore, whatever changes occur in $\mathrm{Tr}\rho \ln \rho$ during the experiment, they cannot be taken to compensate for $(1/T) \int pdV$ since the latter is null throughout the experiment. (Hemmo and Shenker, 2006, p.162)

3.1.4 Discussion of Hemmo and Shenker's (2006) Argument

This section will discuss the results presented above. I will identify two errors in the argument. The first concerns a wrong calculation of the von Neumann entropy during the Step 4 location measurement. The second error regards the unitary reset of the measurement apparatus. If such a unitary reset were indeed possible, then the (orthodox and probabilistic) second law could be violated on a regular basis.

Redundancy of the Step 4 Location Measurement

I will begin the discussion of the above with some general observations, in order to provide some more clarity. For this, we recall that according to Hemmo and Shenker, the only difference between their and von Neumann's original thought experiment is that a *further* measurement, a location measurement (Step 4), is needed to determine the molecule's location prior to the compression stage. For gases at the thermodynamic limit, this measurement becomes redundant, since the amount of molecules on each side of the box becomes proportional to their respective occupying volume. Not so for single molecules, for which, before the empty side of the box can be compressed (with probability one), require a location measurement in order to determine which side the particle is on.

Contrary to H&S's assertions, however, the location measurement during Step 4 is *not* needed. A spin z measurement already took place at Step 2 and the outcome of this measurement will be fully correlated with the position of the particle after the separation in Step 3. And so instead of introducing yet another auxiliary system that performs a location measurement on the particle, it would have been sufficient to read out the measurement result of the spin z measurement.

In the case of collapse, for example, the particle will have already collapsed into a spin eigenstate during the Step 2 measurement. The correlations established during the location measurement will thereby all be classical and reading out the spin-measurement result is sufficient to predict the particle's location after the separation. In the case of no-collapse interpretations and ignoring the system's spatial degrees of freedom and decoherence, system and (spin-)measurement apparatus become entangled during Step 2:

$$|\Psi\rangle^{(2)} = \frac{1}{\sqrt{2}} \left(|+_z\rangle_P |+\rangle_M + |-_z\rangle_P |-\rangle_M \right) \tag{3.6}$$

During the separation in Step 3, whether or not the system is to be found the state of the *box*, namely whether it contains a molecule in the left compartment $|L\rangle_B$ or in the right compartment $|R\rangle_B$ then also becomes entangled with the spin degrees of freedom of the molecule, which means that the state of the particle-apparatus-box system is

$$|\Psi\rangle^{(3)} = \frac{1}{\sqrt{2}} \left(|+_z\rangle_P |+\rangle_M |L\rangle_B + |-_z\rangle_P |-\rangle_M |R\rangle_B \right). \tag{3.7}$$

Therefore, for both collapse and no-collapse cases it is in fact sufficient to read out the measurement result of the Step 2 spin measurement in order to determine the

location of the particle after the separation process.

Having two instead of one measurement(s) would not be much of a problem, if it weren't the case that for H&S, the two measurements have different consequences for the von Neumann entropy. This is the first inconsistency in their argument: whereas H&S agree that after the spin z measurement at Step 2 the *post spin measurement* density matrix of the particle is given by

$$\rho^{(2,red)} = \frac{1}{2}\left(|+_z\rangle\langle+_z|_P + |-_z\rangle\langle-_z|_P\right)\rho(L), \tag{3.8}$$

they do not apply the same reasoning to the *post location measurement* state of the system at Step 4. Instead, they use a 'collapsed' density matrix to calculate the von Neumann entropy:

$$\rho^{(4,+)} = |+_z\rangle\,\langle+_z|_P\,\rho(L)_M \quad \text{or} \quad \rho^{(4,-)} = |-_z\rangle\,\langle-_z|_P\,\rho(R). \tag{3.9}$$

In the first case, the (spin) measurement has therefore increased the entropy, whereas in the second case the (location) measurement has effectively reduced it. What is going on?

Let me first try to assemble what the authors themselves could have had in mind. In von Neumann's original argument, the Step 2 spin measurement is *non-selective*[7], which means that even if the system has *de facto* collapsed into one of its eigenstates, an external agent[8] would not be able to determine into which state the system

[7]Or rather should have been, given that von Neumann himself begins his argument with a spin mixture. This however does not matter for conceptual purposes.

[8]Some words of clarification regarding my use of the term 'agent': an agent does or course not need to be a human being but can be anything that is able to measure and react to the measurement outcome accordingly. For this reason I will use the term 'agent' interchangeably with the term 'measurement apparatus' or even 'memory cell', implying that even a simple binary system can

has collapsed and would therefore describe the system by a density matrix $\rho^{(2,red)}$. The system would be in a so-called proper mixture, meaning that it is possible to understand ρ as representing a probability distribution over pure states[9]. The von Neumann entropy of the system at Step 2 has thereby increased compared to its previous state, in agreement with what von Neumann considers being the irreversibility of a 'Prozess 1'.

The Step 4 location measurement on the other hand is *selective* — it establishes correlations between the agent who performs the measurement[10] and the system. These correlations then allow the agent to perform further operations on the system, such as the Step 5 compression of the box. For H&S, the von Neumann entropy of the system at Step 4 has therefore decreased from Step 3.

Since H&S want to include the non-selective spin-measurement at Step 2, the Step 4 location measurement is indeed a necessary requirement for the single particle case, given that the compression is not being allowed to be conditional on the outcome of the Step 2 measurement. Without the selective measurement, the work-free compression against vacuum could not take place. The limiting case of infinite particles (and in fact von Neumann's original account) does not require this selective measurement since the amount of particles within each chamber of the box becomes equal.

The problem with including a second measurement on the system is that this second measurement also introduces a second measurement apparatus. It will be shown shortly that H&S's conclusion is based on an erroneous calculation of the

serve as an 'agent'. Using the term 'agent' in this way therefore does not imply subjectivity of any kind.

[9]In the case of no-collapse interpretations and ignoring decoherence, the agent is not yet entangled with the system. She only becomes entangled once she performs the Step 4 location measurement.

[10]More precisely the location measurement apparatus, but I will take those two to be synonymous for the time being.

von Neumann entropy when the system is correlated to this second measurement apparatus. Before elaborating on this point however, I would like to discuss another shortcoming of their argument. And, given that it allows us to arbitrarily violate the second law, a severe one moreover.

Violation of the Second Law

H&S notably claim that the thermodynamic entropy change is zero throughout the whole cycle and in particular, that at the end of the cycle "[...] the measuring device [can be returned to its initial ready state] unitarily." (Hemmo and Shenker, 2006, p.161), and hence without any heat cost.

To see why this better not be the case, we consider the consequences of H&S's assertions and assume that it is indeed possible to reset the measurement apparatus without a compensating heat transfer into the environment (against Landauer's principle). We may then construct a slightly amended version of H&S's proposed cycle. No dramatic changes are made, the only thing that changes is Step 5, which instead of being a compression we turn into an isothermal expansion: that is, instead of compressing the empty side against the vacuum, we let the particle push against the partition in a quasi-static, isothermal fashion. Given that the position of the particle is 'known' as a result of the location measurement, it is possible to attach a weight to the partition, thereby extracting $kT \ln 2$ units of work from the system during the expansion, while the according amount of heat is delivered from the heat reservoir.

After the work extraction, the measurement apparatus is brought back to its initial state (which according to H&S can be done for free). The partition is then re-inserted into the original system (for free), the position of the particle measured again (for

free) and the above process is repeated, thereby extracting arbitrarily large amounts of work from this one-particle engine with the sole effect being that heat is extracted from a single reservoir. But this is a direct violation of the Kelvin-Planck statement of the second law (Planck, 1991). Something must have gone wrong!

The problem lies in the statement that the measurement device can be returned to its initial ready-state unitarily. This does not work and mathematically, it can be easily seen why not: since the measurement device after the measurement is in one of the two mutually exclusive states $|-\rangle_M$ or $|+\rangle_M$, there exist no unitary operator that reliably maps the memory cell back to its initial, ready state $\{|-\rangle_M, |+\rangle_M\} \mapsto |ready\rangle_M$.

As discussed in the previous section about the Szilard engine: in order to reset the measurement device unitarily, it is necessary to record beforehand, in which one of the two states the device is in. This however would require a further measurement of yet another measuring device on the first. But then one would want to reset this second measuring device unitarily, too, for which a third device would be needed and so on. Eventually, one would run out of resources and a Landauer erasure would be unavoidable. A unitary reset of the memory cell is hence not possible.

Once the heat cost associated with the resetting operation is taken into account, the complete cycle ceases to be thermodynamically entropy neutral, as the entropy in the environment will have increased in the last step.

Note that H&S cannot get around conceding to this point if they want their argument to persist. They (at least for the purpose of their paper) agree with von Neumann that

> [...] in the sense of phenomenological thermodynamics, each conceivable process constitutes valid evidence, provided that it does not conflict with the two fundamental laws of thermodynamics." (von Neumann, 1996, p.192)[11].

[11] [...] im Sinne der phänomenologischen Thermodynamik ist jeder denkbare Prozess beweiskräftig,

Allowing for a dissipationless reset of the memory cell could in principle violate the second law (even the probabilistic one), therefore clearly conflicting with thermodynamics. In a more generalised account, Ladyman et al. (2008) show that in fact *any* violation of Landauer's principle leads to a violation of the second law. Bad enough as this is, it becomes even worse in the face of their article's objective, which is to show that the thermodynamic entropy differs in behaviour from von Neumann's entropy. But if the cycle is not even thermodynamically consistent (i.e. disobeys the laws), then how could it possibly be used to show that von Neumann entropy is not thermodynamic entropy?

Which Entropy?

Notwithstanding the above criticism, H&S's main point that the von Neumann entropy (as opposed to the thermodynamic entropy) during the Step 4 location measurement *decreases*, and that this gives us reason to reject their conceptual equivalence, still stands, or seems to.

> As a result of the location measurement, the von Neumann entropy decreases back to its original value. (Hemmo and Shenker, 2006, p.163)

In this section I will discuss this claim and in particular I will show that:

(i) The physically relevant entity is the joint entropy of system *and* measurement apparatus. It remains the same.

(ii) It is the system's so-called conditional entropy that decreases during the measurement, not the system's marginal von Neumann entropy.

wenn er die beiden Hauptsätze nicht verletzt.

I will now begin with a justification of the two claims. In phenomenological thermodynamics, the joint entropy of two systems is always the sum of their respective entropies. The von Neumann (in the classical case the Gibbs-) entropy on the other hand is generally subadditive and additive only in the absence of correlations between two systems:

$$H(S, M) \leq H(S) + H(M), \tag{3.10}$$

where S stands for 'system' and M stands for 'measurement apparatus' or 'memory cell'. In the concrete case of the Step 4 location measurement, we can model the measurement apparatus M as a box containing a single molecule and divided by a partition. It can then be in one of two mutually exclusive states, corresponding to the position of the molecule, left (l) or right (r). We assume it needs to be in a 'ready'-state before the measurement, which we chose to be l. For simplicity, we assign an entropy of zero to it. If we consider the case of collapse, at the time of the measurement the system will have already collapsed into a spin eigenstate. The correlations between the location of the system and the measurement apparatus are then all essentially classical and the von Neumann entropy before the measurement can be rewritten as:

$$H(S) = -k_B \sum_{s=l,r} p(s) \log p(s), \tag{3.11}$$

where $p(s)$ is the probability of the system being in the left or right chamber of the box.

Before the Step 4 location measurement, system and measurement apparatus are not correlated, and their joint entropy is given by $H_3(S, M) = H_3(S) + H_3(M)$, where

3 is taken to denote 'Stage 3', or, in other words, 'before the Step 4 measurement'. During the measurement, the memory cell will align itself with the position of the particle and the two systems become correlated. The joint entropy now cannot be expressed anymore as the sum of the individual entropies and instead becomes

$$H_4(S, M) = H_4(S|M) + H_4(M) \leq H_4(S) + H_4(M), \tag{3.12}$$

with $H(S|M)$ being the so-called *conditional entropy*, which quantifies how much S is correlated with M and which is given by

$$H(S|M) = -k_B \sum_{s,m} p(s, m) \ln p(s|m) \tag{3.13}$$

$$= -k_B \sum_{m} p(m) \sum_{s} p(s|m) \ln p(s|m) \tag{3.14}$$

$$= k_B \sum_{m} p(m) H(s). \tag{3.15}$$

$p(s)$ and $p(m)$ are the probabilities that system and memory cell are found in macrostate $s = l_s, r_s$ or $m = l_m, r_m$ respectively, $p(m, s)$ is their joint probability and $p(s|m)$ the conditional probability, with $H(s) = -k_B \sum_s p(s|m) \ln p(s|m)$. The conditional entropy is non-negative and is maximal when system and measurement apparatus are uncorrelated, , $0 \leq H(S|M) \leq H(S)$, in which case Equation (3.12) reduces to Equation (3.10). It is often considered to be the entropy relative to an agent (in this case the measurement apparatus).

Let us now go back to H&S's claim that the von Neumann entropy of the system decreases during the location measurement. Does it? The answer is no. What decreases, however, is the *conditional entropy* relative to the measurement apparatus:

$$H_3(S|M) \geq H_4(S|M). \tag{3.16}$$

It reduces to zero, because system and measurement apparatus become perfectly correlated during the measurement. And so when H&S claim that the system's entropy has decreased, what they *mean* is that the system's conditional entropy has decreased. But the conditional entropy is distinct from the marginal entropy. Re-writing the joint entropy of system and measurement apparatus demonstrates this:

$$H_4(S, M) = H_4(S|M) + H_4(M) = H_4(M|S) + H_4(S). \tag{3.17}$$

As opposed to phenomenological thermodynamics, which treats systems as black boxes, (classical and quantum) statistical mechanics is able to detect correlations between subsystems, allowing us to mathematically handle the concept of 'measurement' in the first place. If we associate entropy with the potential to (reliably) extract work from a system, then the conditional entropy certainly quantifies this ability to a certain extent: a memory cell endowed with an automaton would now be able to (reliably) extract work from the system by allowing it to isothermally expand into the other half of the box, thereby raising a weight. Contrary to an external agent who is not correlated to the particle location. But this is just the ordinary Maxwell's demon scenario[12] applied to a one-particle setting (Szilard, 1929).

[12]As mentioned before, Maxwell in 1867 introduced the idea of a "very observant and neat-fingered being" (as cited in (Maxwell, 1995)), which was intended to demonstrate that the orthodox second law of thermodynamics could be broken in principle by exploiting fluctuations. In the thought experiment, a box filled with monoatomic gas is divided into two parts by a partition into which a small door is inbuilt. The "being", later called Maxwell's demon, controls every atom that approaches the door and either lets the atom pass or not. Since the gas molecules are subject to a velocity distribution, he can decide to only let the fast molecules pass into the one direction and to only let slow molecules pass into the other direction. By doing so the demon creates a temperature

What becomes important for thermodynamic treatments in such a setting, is the joint entropy of system *and* measurement apparatus, as the joint system (ideally) has no correlations with the outside and can thus be treated as a thermodynamic black box. And it turns out that the behaviour of the thermodynamic entropy of the joint system, is exactly mirrored by the behaviour of the von Neumann entropy: the *joint* entropy of system and measurement apparatus does not change during the location measurement, but remains the same:

$$H_3(S, M) = H_4(S, M). \tag{3.18}$$

And so, to summarise the above: all that changes during the location measurement is the conditional entropy, but neither the joint entropy of the system nor the marginal entropy $H(S)$. Furthermore, the joint entropy, just as the thermodynamic entropy of the joint system, remain the same during the location measurement.

No Collapse Scenarios

Let us now consider the case of no collapse scenarios. In no collapse scenarios, following the measurement in Step 4, the measurement apparatus and the location degree of freedom become entangled. The reduced density matrix after tracing out the decohering environment therefore is given by an inproper mixture due to the neglect of the correlations with the environment. After the Step 4 measurement, the density matrix of the combined system and measurement apparatus is given by

$$\rho^{(4,P+M)} = \frac{1}{2}\left(|+_z\rangle\langle+_z|_P\,\rho(L)_P\,|L\rangle\langle L|_M + |-_z\rangle\langle-_z|_P\,\rho(R)_P\,|R\rangle\langle R|_M\right), \tag{3.19}$$

gradient, allowing him to violate the second law.

where now $|L\rangle_M$ and $|R\rangle_M$ represent the states of the measurement apparatus. The correlations between the measurement apparatus and the system are of a classical nature and so also in the absence of collapse, the von Neumann entropy of system and apparatus has not changed during the Step 4 measurement.

3.1.5 What could have been von Neumann's Response?

After the above discussion, it may be interesting to explore what von Neumann himself might have thought of H&S's claim that his reasoning does not establish a conceptual link between von Neumann entropy and thermodynamic entropy. Chances are he would not have accepted their conclusion. This section elaborates why.

In his original setup, von Neumann introduced a gas consisting of individual quantum systems, locked up in boxes and placed in a further, giant box. This setup was first proposed by Einstein (1914). Von Neumann has in mind a representative, imaginary statistical (but finite) *ensemble* (*Gesamtheit*[13]). For him, the density operator (which he calls the 'statistical operator') can only relate to such a *Gesamtheit*. This means that even in the case of an individual quantum system, von Neumann's argument would remain unchanged: the density operator of this individual quantum system would *still* relate to an ensemble of systems and a system containing a single particle would therefore *still* be modeled as an N particle ensemble. The statistical representations of a) a system containing a single particle, and b) a system containing many particles, are therefore identical. And so chances are that von Neumann would have rejected H&S's discussion on the basis of a misunderstanding of the statistical operator itself.

[13]In footnote 156, von Neumann explains that he acquired the notion of an ensemble by studying von Mises (1936), suggesting that von Neumann has in mind a frequentist understanding of probabilities.

The statistical operator, hence equally the von Neumann entropy, are not concerned with the interacting constituents of a single (even if very complicated) mechanical system, but rather with such a *Gesamtheit* of many, independent and non-interacting individual systems (von Neumann, 1996, p.191). This however does not mean that von Neumann denies that one can meaningfully apply thermodynamics to individual particles (von Neumann, 1996, p.212). Quite the contrary, he explicitly considers the case of a single particle in a box in order to demonstrate that whether or not an external agent can extract work from the particle, depends on the agent's state of knowledge about the position of the particle. If the observer knows for a fact that the particle is located in the left hand side of the box (partitions notably do not need to be present in the box, at this stage), the entropy of the system is less than if the observer is completely ignorant about the whereabouts of the particle.

One may be inclined to accuse von Neumann of inconsistency, given that on the one hand entropy appears as a statistical concept applicable only to a *Gesamtheit*, but on the other hand it is possible to assign a thermodynamic entropy to a system containing only one particle in a box. This inconsistency disappears once we take entropy as quantifying an observer's epistemic capacities. Von Neumann explicitly writes about 'lack of knowledge' about the position of the single particle. In the quantum case, the system being in a mixed state equally quantifies ignorance, but this time ignorance about which pure state the system is in.

Von Neumann writes:

> The temporal variations of the entropy are due to the fact that the observer does not know everything, or rather that he cannot determine (measure) everything, that is in principle measurable. (von Neumann, 1996, p.213)[14]

[14] "Die zeitlichen Variationen der Entropie rühren also daher, dass der Beobachter nicht alles weiss, bzw. dass er nicht alles ermitteln (messen) kann, was prinzipiell messbar ist."

The problems with such ignorance interpretations are manifold. For one, we need to distinguish between proper and improper mixtures. In case the mixtures are improper, that is in case they arise from tracing out the degrees of freedom of an entangled subsystem, the system cannot be thought of as being in one or the other pure state with a given probability and we cannot use the density matrix as a measure of ignorance. Von Neumann however does not consider this case.

In the case of a proper mixture, the density matrix actually can be given an ignorance interpretation over pure states, which is what von Neumann does. However, since the decomposition of the density matrix is highly non-unique, i.e. since there are many different ensembles with different eigenstates and different probabilities that give rise to one and the same density matrix, the density matrix loses its appeal as being an ignorance measure over pure states here as well. In the particular case of von Neumann's argument we may actually get away with this latter version in the case of a collapse interpretation, since the *Gesamtheit* has been *prepared* to be composed of a specific set of eigenvectors.

To conclude and summarise the above: within the framework of von Neumann's *Gesamtheit* interpretation, his argument is immune to the criticism by H&S, even though such an interpretation ought to be rejected on independent grounds.

3.1.6 Conclusion

This section took a closer look at von Neumann's historical argument in favour of the von Neumann entropy. It was shown that the criticism by Shenker (1999) and Hemmo and Shenker (2006) is flawed because a) their reasoning allows for a perpetuum mobile and b) the alleged decrease in entropy during the Step 4 location measurement is in fact a decrease in conditional entropy. In particular, I showed that

the relevant entropy, *joint* entropy of system and measurement apparatus, remains unchanged during the location measurement.

In the next section, I will continue to investigate thermodynamics in the quantum realm by showing how thermodynamic behaviour arises from quantum mechanics, given a set of assumptions. I will then compare this approach to what has become known as 'quantum resource theories of thermodynamics', which are structurally very similar to the previously discussed means-relative approach to thermodynamics.

3.2 Quantum Thermodynamics

One, if not *the* paradigmatic example of inter-theoretic reduction is the reduction of thermodynamics to statistical mechanics. Still, whether or not this reduction is at all successful remains an issue of controversy within the philosophy community (Sklar, 1999; Callender, 1999; Dizadji-Bahmani et al., 2010). One of the various difficulties in this regard is to decide which statistical mechanical framework thermodynamics ought to reduce to in the first place. In the literature, the two main contestants are Gibbsian and Boltzmannian statistical mechanics, both of which naturally bring advantages and disadvantages to the task at hand (as was seen in Chapter 1). Despite their very different ontologies, the two frameworks nevertheless have one thing in common: they are both based on classical mechanics. Given that at its chore, the world is quantum and not classical, it is surprising that the debate still focusses so heavily on which of these conceptually quite outdated theories provides us with the right statistical generalisation of the thermodynamic entropy[15].

In this section, I want to fill the gap between the debate around the correct statistical

[15] An exception is Wallace (2013a), whose account of statistical mechanics is based on quantum mechanics.

mechanical generalisation of thermodynamic concepts, and the fact that the world is fundamentally quantum. The inclusion of quantum mechanical generalisations into the philosophical discourse is of relevance particularly with regard to the newly emerging field of *Quantum Thermodynamics* (see (Brandão et al., 2013; Rio et al., 2011; Horodecki and Oppenheim, 2013; Skrzypczyk et al., 2014; Brandão et al., 2015; Gour et al., 2015) but in particular Goold et al. (2016) and references therein). This rapidly expanding branch of contemporary physics, which also comprises the so-called *Quantum Resource Theory of Thermodynamics*, discussed in more detail in Section 3.2.6, offers new insights and new challenges for philosophers interested in inter-theoretic reduction and the foundations of statistical mechanics. The main goal of this section is to provide the reader with a quantum statistical mechanical generalisation of thermodynamics, based on work by Maroney (2007)[16], and to explore its philosophical implications. It will be shown how we can understand the familiar thermodynamic entities such as work, heat and entropy within a quantum mechanical setting, based on few assumptions, mostly from phenomenological thermodynamics. Doing so will dissolve some of the worries that are typically associated with the notion of entropy in the classical Gibbsian approach. It should be noted that none of the results hinges on any particular interpretation of phenomenological thermodynamics, but adopting an operational approach (Wallace, 2014; Myrvold, 2011; Jaynes, 1965) certainly helps with providing some relevant intuitions. A further advantage of such an operational understanding of thermodynamics is that dealing with finite and even single particle systems becomes conceptually less problematic, provided infinite heat reservoirs exist. Given that the formalism itself is silent about the size of the system, this is a welcome addition.

[16]And draws on insights by (Gibbs, 1878; Szilard, 1925; Tolman, 1938; Wehrl, 1978; Partovi, 1989).

The main body of this section is structured as follows: I will begin with an outline of phenomenological thermodynamics, followed by a few brief words about the relation between thermodynamics and statistical mechanics. After explicitly stating the assumptions that will underly the subsequent discussion, I will then explore how heat, work and entropy derive from quantum statistical mechanics. In the last part, the quantum resource theoretic approach to thermodynamics is introduced and compared with the present proposal.

3.2.1 On the Relation between Thermodynamics and Statistical Mechanics

A large amount of literature in the foundations of statistical mechanics deals with the question of whether and how thermodynamics reduces to it (for example Sklar (1999); Callender (1999)). While some concepts such as energy and volume are well familiar from other theories, those of temperature and entropy are unique to thermodynamics. And so the debate focusses mainly on whether there exist unique statistical mechanical counterparts to these quantities, and if so, what they are.

In the foundations of statistical mechanics, the two main candidates normally considered as being the generalisation of the thermodynamic entropy are the Boltzmann and the Gibbs entropy, already discussed in Chapter 1. To recall their basic attributes: the former is given by the phase-space volume of the macrostate the thermodynamic system is in at a given time, $S_B = k_B \log \Omega$. The latter by a functional of a probability distribution over phase space, $S_G = -\int k_B p(x) \log p(x) dx$. In many Boltzmannian accounts of statistical mechanics, entropy is understood as being a property of the system's microstate and therefore a property of the thermodynamic system's momentary physical state (see for example (Goldstein, 2001; Lebowitz, 1993)). In

the Gibbsian case, however, entropy is a property of the probability distribution over the system's possible microstates, and the difficulties lie in justifying the use of these probabilities. There is on the one hand the question of whether the use of probabilities prevents us from understanding entropy as a non-epistemic quantity, and on the other hand the question of whether there is tension between the use of (even objective) probabilities and entropy being a property of the thermodynamic system. In the latter case it should be noted that 'being a property of' does not imply that the property must be intrinsic. It could equally be an extrinsic or relational property.

Given that the underlying microdynamics is deterministic, it is often thought that the only feasible way to understand the probabilities is as measures of ignorance (Jaynes, 1957; Albert, 2003) (they take different views on whether this is a good thing) over microstates. If this were true, entropy would equally merely quantify the amount of knowledge an agent has about the system's actual microstate, conflicting with the idea that entropy (and hence the thermodynamic state itself) characterises an objective property of the state a system is in[17].

Before deriving the thermodynamic behaviour of quantum systems, however, I will outline a set of assumptions that may also serve as a road map for the following sections. The first and most basic assumption is that

(i) Thermodynamics is about possible and impossible processes of thermodynamic systems between equilibrium states. These processes are characterised by a set of thermodynamic variables and a set of operations.

This assumptions just explicitly states what was already said before in the first chapter

[17] I take 'being a property of a system' and 'being a property of the state the system is in' to be the same.

of this thesis , with an emphasis on the operational aspect. By 'thermodynamic systems' I mean a systems that can be described by a set of thermodynamic variables, such as volume, temperature, entropy, and so on. The equilibrium states are defined relative to a set of conserved thermodynamic quantities and external constraints on the system[18].

The next assumption is that

(ii) It is meaningful to identify thermodynamic entities with expectation values.

This may be more controversial, given that expectation values introduce the notorious problem of the nature of probabilities which already cause trouble in the classical Gibbsian statistical approach. Another problem is that expectation values may not at all be useful for making predictions about the actual outcome of an experiment. In fact the expectation value may not even be among the *possible* outcomes of an experiment. The first issue will be shown to become less pressing in the quantum context. In particular it becomes easier to avoid reference to uncertainty, or lack of information. The second issue is not as easily resolved, and whether or not expectation values can be taken as representing our well known thermodynamic quantities depends very much on whether one is willing to re-interpret the laws of thermodynamics as referring to average quantities[19] or not. The advantages gained by such a re-interpretation are big: not only can we have a sound physical underpinning to the laws of thermodynamics if heat, work and entropy refer to expectation values. The range of applicability of thermodynamics is also extended significantly. Small-scale systems in which fluctuations are relevant are just as much

[18] An additional assumption is that the system approach equilibrium and remain there, as stated by the minus first law of thermodynamics Brown and Uffink (2001).

[19] By this I do not mean that "Thermodynamics is true on average", but rather that "Re-interpreted thermodynamics is true".

subject to these re-interpreted laws as 'ordinary' thermodynamic systems. This is as opposed to the orthodox law which, as already discussed, is violated on a regular basis.

The next important assumption is:

(iii) Heat reservoirs exist.

I have already given Fermi's definition of a heat reservoir in Chapter 1. To recap, a heat reservoir can only exchange heat, not work, with its surroundings and is uniquely characterised by its temperature. In particular it is assumed to be sufficiently large such that it temperatures changes only negligibly when it exchanges heat with other bodies. This assumption is borrowed from phenomenological thermodynamics and it turns out that it is incredibly useful to motivate the representation of thermal states by the canonical states. Once this has been established, everything else, including entropy and the second law, follows.

With the above assumptions at hand, we can now derive all of the relevant thermodynamic equations, as will be shown now.

3.2.2 Heat and Work in Quantum Statistical Mechanics

Central to thermodynamics is the distinction between heat and work. This section will show that this distinction between two types of energy transfer can be naturally extended to quantum mechanics.

An isolated quantum system is described by its state $\rho : \mathcal{H} \to \mathcal{H}$, a bounded, positive trace-class operator with trace one, where $\mathcal{H}$ denotes the system's complex Hilbert space. Its evolution is given by the *Liouville-von Neumann equation:*

$$i\hbar\frac{\partial\rho}{\partial t} = [H, \rho]\,, \tag{3.20}$$

where the hermitian operator H is the system's Hamiltonian. The above equation resembles the classical Liouville equation which describes the evolution of a probability distribution over phase space. In the classical case, the Liouville equation conserves phase space volume. Here in the quantum case, the it is the eigenvalues of the density operator that remain constant throughout the Hamiltonian evolution.

The (mean) energy of a quantum state ρ is given by the expectation value of its Hamiltonian, $\langle H\rangle_\rho = Tr\,[H\rho]$. We now consider an isolated quantum system, whose evolution solely depends on H. If the Hamiltonian is varied (this may be done by the manipulation of an external field), the rate of change of the system's mean energy is given by

$$\frac{\partial\langle H\rangle_\rho}{\partial t} = \left\langle\frac{\partial H}{\partial t}\right\rangle_\rho, \tag{3.21}$$

If we now demand that the Hamiltonian is varied[20] in a cyclic fashion, namely $H(t_0) = H(t_1) = H_0$, then integrating Equation (3.21) leads to

$$W(\rho) = Tr\,[H_0(\rho(t_1) - \rho(t_0))]\,. \tag{3.22}$$

This describes the amount of energy that can be extracted from a quantum system by a cyclic variation of the Hamiltonian. Given that the system is not in direct contact with any other system during this operation, we may identify the above with the average *work* $W(\rho)$ that is performed by the system. It is given by the difference

[20]We take the Hamiltonian to be a function of some parameters (x, y, z) that are varying in time, so that $H = \sum_n E_n(x, y, z)\,|E_n(x, y, z)\rangle\,\langle E_n(x, y, z)|$. See Maroney (2007) for more details.

in energy expectation values. A discussion on the interpretation and legitimacy of expectation values in quantum thermodynamics will follow later.

Equation (3.21) only holds for isolated systems. If the system is allowed to interact with another system, however, the energy flux due to the interaction needs to be taken into account. To show this, we begin by considering two quantum systems in a product state $\rho(t) = \rho_S(t) \otimes \rho_E(t)$, where S stands for 'system' and E for 'environment'.[21] The states of the subsystems are given by their reduced density matrices $\rho_S(t) = Tr_E[\rho(t)]$ and $\rho_E(t) = Tr_S[\rho(t)]$. The combined system evolves under the Hamiltonian $H = H_S + H_E + V_{SE}$, where $H_S \in \mathcal{O}(\mathcal{H}_S)$, $H_E \in \mathcal{O}(\mathcal{H}_E)$ and the interaction potential $V_{SE} \in \mathcal{O}(\mathcal{H}_S \otimes \mathcal{H}_E)$. For simplicity we consider the case for which the interaction potential is switched on at time t_0 and the Hamiltonian is then varied. We can then describe the evolution of the total system's energy as

$$\frac{\partial \langle H \rangle_\rho}{\partial t} = \frac{\partial \langle H_S \rangle_{\rho_S}}{\partial t} + \frac{\partial \langle H_E \rangle_{\rho_E}}{\partial t} + \frac{\partial \langle V_{SE} \rangle_\rho}{\partial t}. \tag{3.23}$$

The evolutions of the individual subsystem's energies are given by

$$i\hbar \frac{\partial \langle H_S \rangle_{\rho_S}}{\partial t} = i\hbar \langle \frac{\partial H_S}{\partial t} \rangle_{\rho_S} + \langle [H_S, V_{SE}] \rangle_\rho, \tag{3.24}$$

$$i\hbar \frac{\partial \langle H_E \rangle_{\rho_E}}{\partial t} = i\hbar \langle \frac{\partial H_E}{\partial t} \rangle_{\rho_E} + \langle [H_E, V_{SE}] \rangle_\rho, \tag{3.25}$$

$$i\hbar \frac{\partial \langle V_{SE} \rangle_\rho}{\partial t} = i\hbar \langle \frac{\partial V_{SE}}{\partial t} \rangle_\rho - \langle [H_S, V_{SE}] \rangle_\rho - \langle [H_E, V_{SE}] \rangle_\rho. \tag{3.26}$$

As opposed to Equation (3.21), Equation (3.24) contains an extra term $\langle [H_S, V_{SE}] \rangle_\rho$

[21] This initial condition of the systems being in a product state is similar to Boltzmann's *Stosszahlansatz* in that it takes the systems to be uncorrelated at the beginning.

that contributes to the change in energy and that arises from the interaction with E. The evolution of subsystem S is therefore not closed, since the second term on the right hand side of Equation (3.24) still depends on the total state $\rho(t)$.

At this point it is useful to make some simplifying assumptions. If the interaction potential V_{SE} is taken to be constant in time and furthermore, if we demand a finite interaction between the two systems such that the systems are only interacting at time $t_0 < t < t_1$, then $\langle V_{SE} \rangle_{\rho(t_0)} = \langle V_{SE} \rangle_{\rho(t_1)} = 0$. Integrating Equation (3.26) and re-arranging the terms then leads to

$$\int_{t_0}^{t_1} \langle [H_S, V_{SE}] \rangle_{\rho(t)} \, dt = - \int_{t_0}^{t_1} \langle [H_E, V_{SE}] \rangle_{\rho(t)} \, dt. \tag{3.27}$$

The above equation states that the interaction-induced energy increase in system S is equal to the energy decrease of system E. Labelling the left hand side Q_S and the right hand side $-Q_E$, the above equation takes the more familiar form:

$$Q = Q_S = -Q_E. \tag{3.28}$$

After integration, equations (3.24-3.26) can now be rewritten in a way that resembles the thermodynamic first law:

$$\Delta U_S = W_S + Q \tag{3.29}$$

$$\Delta U_E = W_E - Q, \tag{3.30}$$

where $\Delta U_i = \langle H_i(t_1) \rangle_{\rho_i(t_1)} - \langle H_i(t_0) \rangle_{\rho_i(t_0)}$ is just the change of energy for system i. In case the Hamiltonian H_E of E is time-independent, W_E vanishes. This is desirable

when we want E to be a heat bath, which by definition is only able to interact with other system via heat exchange.

The way it was presented, Q played the role of the quantum mechanical analogue to the thermodynamic notion of heat. The notion of heat in thermodynamics however is much richer as the Q presented above, not least due to its relation to absolute temperature and entropy.

3.2.3 Thermal States and Heat Baths

The previous section showed that quantum mechanics provides a natural way to distinguish between two types of energy transfer, Q and W, and introduced an equation that resembled the phenomenological first law. A lot of the richness and the power of thermodynamics however derive from the second law, which defines an absolute temperature and limits the thermodynamic processes to those that are entropy non-decreasing. By doing so, it also restricts the amount of work one can reliably extract from a thermodynamic system.

The Canonical State

In classical thermodynamics, prior to the notion of entropy is the notion of thermal equilibrium (Brown and Uffink, 2001). Any statistical mechanical generalisation therefore needs to be able to explain the approach to and the remaining in thermal equilibrium. In Boltzmannian statistical mechanics, equilibrium is associated with the system being in the equilibrium macrostate. In Gibbsian statistical mechanics on the other hand, equilibrium is associated with a stationary probability distribution over phase space - for thermal systems, this is the canonical distribution. In the quantum case, the equilibrium state is similarly given by the canonical state

$$\rho_\beta = \frac{e^{-\beta H}}{Tr\left[e^{-\beta H}\right]}, \tag{3.31}$$

with $\beta = 1/k_B T$. Derivations of ρ_β as the unique thermal distribution face several problems. The majority of arguments at some point or other needs to make reference to an *a priori* probability distribution (which itself will need justification). In the quantum case, this is associated with the maximally mixed state[22].

Following Maroney (2007), I want to motivate the canonical state in a slightly different way, namely based on phenomenological considerations: The second law as encountered in Section 1 applies to closed systems which consist of i) bodies that undergo cyclic processes, ii) a weight that can be lowered and raised and iii) numerous heat reservoirs at different temperatures (Szilard, 1925). The temperature T in the original formulations of the second law and in the Clausius inequality refers to the temperature of the heat baths, as was already mentioned. Heat baths therefore play an essential role in determining the question of what it means for a system to have a temperature T and what it means for it to be in a thermal state. The first task of this section will be to find an appropriate quantum model for a heat reservoir and to show that the canonical state is the only feasible candidate as representing the thermal states that constitute the heat baths. After this, everything else smoothly falls into place.

A heat bath is a system, which, by definition, can only exchange heat but not work with its environment. The first task will then be to find quantum states that have

[22]Popescu et al. (2005) and Gemmer et al. (2009) have found a way around the *a priori* equal probability distribution, due to some constraints. The Hilbert space of the universe is comprised of the system and a larger environment and the universe itself is postulated to be in a pure state. Popescu et al. (2005) then show that for nearly all pure states the universe could be in, the system as a subsystem of the universe behaves *as if* the universe were in a maximally mixed state. From there it follows that the system, being sufficiently small compared to the overall universe, is approximately described by the canonical distribution for almost every pure state.

exactly these attributes, to find systems that can interact, but from which no work can be extracted themselves, namely by means of unitary transformations. Such states are called *passive* states (Pusz and Woronowicz, 1978; Lenard, 1978; Sewell, 1980). Any system that is governed by a Hamiltonian that is bounded from below (and most physical Hamiltonians are), will be able to be in such a passive state. This also restricts the amount of extractable energy from a system.

Following Equation (3.22), the amount of work extracted from an isolated system is maximised if the energy of the final state $\rho(t_1) = U\rho(t_0)U^\dagger$ is minimized. This is the case if $\rho(t_1)$ is diagonalised in the energy eigenbasis

$$\rho(t_1) = \sum_i p_i \left|E_i\right\rangle \left\langle E_i\right| \tag{3.32}$$

such that

$$p_i \geq p_j \leftrightarrow E_i \leq E_j \tag{3.33}$$

If two systems ρ_1 and ρ_2 are both passive, their joint system $\rho_{12} = \rho_1 \otimes \rho_2$ however need not be. For n systems to be jointly passive[23], it is a sufficient condition that for any pair of energy levels (i, j),

$$\frac{\ln(p_i/p_j)}{E_j - E_i} = \beta, \tag{3.34}$$

where β is a constant. Re-arranging shows that for n systems to be jointly passive, their joint probabilities are given by

$$p_i = \lambda e^{-\beta E_i}, \tag{3.35}$$

[23]See (Maroney, 2007, p.19-21) for a derivation of joint passivity.

where λ is another constant. But this is just the canonical ensemble. Hence, for the product state of n passive systems to be again a passive system, the canonical ensemble is the unique completely passive ensemble[24].

If we want to model a heat reservoir quantum mechanically, the canonical state seems to be the only feasible candidate, as it is the only uniquely passive state and so it is the only state that guarantees that no work can be extracted from the heat bath, just as required by the phenomenological theory. The heat bath will therefore be modelled as a large assembly of uncorrelated[25] canonical systems.

Thermalisation

The next step is to show that systems that come in contact with the heat bath, and that begin in different (not necessarily thermal) states, eventually thermalise and settle down into a thermal state with parameter β.

To demonstrate that this is indeed the case we first introduce the von Neumann measure:

$$G(\rho) = Tr\left[\rho \ln \rho\right] \tag{3.36}$$

For a fixed energy $\langle H \rangle_\rho = E$, the von Neumann measure is minimised by the canonical state ρ_β, and so $G(\rho_\beta) \leq G(\rho)$, where ρ is just any other state on the same space. The introduction of (3.36) may seem slightly *ad hoc*, but for now $G(\rho)$ is merely used as a mathematical tool.

For a fixed value of β, the canonical state ρ_β also minimises

[24]Note that a system can only be passive if its density matrix commutes with its Hamiltonian, i.e. if the system can be diagonalised in its energy eigenbasis.

[25]Uncorrelated in the sense that the heat bath is not correlated with any other system but also in the sense that the canonical subsystems of the heat bath are not correlated with each other. This is required because we do not want the entanglement between an individual subsystem and the target system to spread over the whole heat bath.

$$G(\rho) + \beta\langle H\rangle_\rho \tag{3.37}$$

This, implies that[26]

$$G(\rho) + \beta\langle H\rangle_\rho \geq G(\rho_\beta) + \beta\langle H\rangle_{\rho_\beta}. \tag{3.38}$$

We now consider two systems, the first one in an arbitrary, the second in a canonical state. They start out in a product state $\rho_S \otimes \rho_{E,\beta}$, where S stands for 'system' and E stands for 'environment'. They are then allowed to interact for a finite period of time t via the Hamiltonian $H = H_S + H_E + V_{SE}$ (the interaction is 'switched on' at time t_0 and then switched off at time t_1) where H_S and H_E act on the respective subspaces of ρ_S and $\rho_{E,\beta}$, and V_{SE} acts on the combined space and in a way that the total energy is conserved, $\langle H\rangle_\rho(t_0) = \langle H\rangle_\rho(t_1)$.

Then for the marginal state of the system $\rho'_S(t) = Tr_E\left(\rho_{SE}(t)\right)$ after some time t we find that the interaction with a thermal state *decreases* the quantity $G(\rho_S)+\beta\langle H_S\rangle_{\rho_S}$, for an arbitrary state ρ:

$$G\left(\rho'_S(t_0)\right) + \beta\langle H_S\rangle_{\rho'_S(t_0)} \geq G\left(\rho'_S(t_1)\right) + \beta\langle H_S\rangle_{\rho_S\prime(t_1)} \tag{3.39}$$

If we now model the heat bath as a large assembly of canonical states as bath-subsystems, all with the same parameter β, then we can describe how a system in an arbitrary initial state evolves into a canonical state with the same parameter β. For this, the system is coupled to a bath-subsystem for a small small amount of

[26]Calculations that explicitly show why ρ_β minimises (3.36) and (3.37) can be found in most textbooks on statistical mechanics, where $G(\rho)$ is identified with negative entropy and $G(\rho)+\beta\langle H\rangle_\rho$ with the free energy.

time (more about this later) and then decoupled again. The system is then again coupled to another bath-subsystem for a short time-period and so forth. With each decoupling the system will have evolved closer to an equilibrium state ρ_∞. If this equilibrium state is the canonical state, $\rho_\infty = \rho_\beta$, we can say that the system *thermalises*. I will furthermore from now on identify β with the thermodynamic temperature, $\beta \equiv 1/k_B T$. In the next section, I will say some more words about heat baths and some problems that may arise.

Real Heat Baths

The above model for thermalisation is based on the assumption of *ideal* heat baths. In order to show thermalisation for concrete, real heat baths, there will always be some approximations involved. The bath-subsystems for example are typically treated as being Markovian — in particular if the heat bath is not infinitely large, which real heat baths never are, we require that the interaction between the bath-subsystems are sufficiently weak. Otherwise a correlation established between the system and one of the bath-subsystems would spread over the bath, potentially establishing correlations between system and another bath's subsystem the system has not encountered yet. *Real* heat baths will always have such correlations. In practice, such formed correlations between system and bath-subsystem will be irretrievably lost due to the size of the heat bath. In fact, even for non-ideal heat baths with strong correlations between the constituents, the correlations relative to the system itself decrease with $1/N$, where N is the number of bath sub-systems (Partovi, 1989). This way, even for the inevitable existence of correlations between subsystems in the heat bath, with a large enough chosen reservoir, they will not influence the target system. The possibility of *recurrence* will however always exist when considering real heat baths.

Curiously enough our heat bath better not be *too* ideal, that is, we don't want our system to interact with the subsystems for an arbitrarily *short* time. The reason for this is, that we can understand the interaction of a system with one of the heat bath's subsystems as a kind of *measurement*. If the time between these measurements becomes too short, the system would be dynamically *frozen*, a phenomenon also known as the Quantum Zeno effect (Erez et al., 2008; Misra and Sudarshan, 1977). Formally, this means that the time derivative of the system's density matrix vanishes immediately after the measurement $\dot{\rho}_S(t) = 0$. The Quantum Zeno effect prevents us from taking the limit of infinitely short time-intervals between interactions, as we could do in the classical case. The maximal frequency of successive measurements without encountering the Quantum Zeno effect mainly depends on the separation of the system's energy levels. Most results on the Quantum Zeno effect in the context of thermodynamics however are based on very simplified models of two-level systems (Schulman and Gaveau, 2006). Curiously enough, if we increase the time intervals between measurements slightly, we may even hit the *Anti*-Zeno effect (Kofman and Kurizki, 2001), which accelerates the process.

3.2.4 Entropy and the Second Law

The above sections introduced the notions of work, heat, thermal states and heat baths. This part will finally introduce the notion of entropy. It turns out that at this point, all the relevant steps for picking out the von Neumann entropy as the correct quantum mechanical realisation of the thermodynamic entropy, have already been made. Re-arranging Equation (3.39) leads to

$$G(\rho'_S(t_0)) - G(\rho'_S(t_1)) \geq \beta \left(\langle H_S \rangle_{\rho'_S(t_1)} - \langle H_S \rangle_{\rho'_S(t_0)} \right). \tag{3.40}$$

From now on I will use the standard expression for the von Neumann entropy $S(\rho) \equiv -G(\rho) = -Tr\rho \ln \rho$ and I will drop the primes above as it should be clear that they refer to the reduced density matrices. The above equation then becomes

$$S\left(\rho_S(t_1)\right) - S\left(\rho_S(t_0)\right) \geq \beta \left(\langle H_S \rangle_{\rho_S(t_1)} - \langle H_S \rangle_{\rho_S(t_0)} \right) \tag{3.41}$$

We now consider the case of two canonical systems ρ_1 and ρ_2 at temperatures β_1 and β_2. As before, the systems are allowed to interact for a finite period of time, but such that the overall energy remains constant. It is then the case that

$$\beta_1 \left(\langle H_1 \rangle_{\rho_1(t_1)} - \langle H_1 \rangle_{\rho_1(t_0)} \right) + \beta_2 \left(\langle H_2 \rangle_{\rho_2(t_1)} - \langle H_2 \rangle_{\rho_2(t_0)} \right) \geq 0. \tag{3.42}$$

Given that the total energy is conserved during the process, the above reduces to

$$\left(\langle H_1 \rangle_{\rho_1(t_1)} - \langle H_1 \rangle_{\rho_1(t_0)} \right) (\beta_1 - \beta_2) \geq 0. \tag{3.43}$$

From the above we can see that if there is a positive flux of energy into the system and the first term is positive, then $\beta_S > \beta_E$, or, differently put, $T_S < T_E$. In words, this means that heat on average has to flow from warm to cold.

The Quantum Clausius Inequality

In phenomenological thermodynamics, the Clausius inequality is derived by having a system interact with a number of ideal heat baths at temperature T_i. Taking $\langle H_i \rangle_{\rho_i(t_1)} - \langle H_i \rangle_{\rho_i(t_0)} = Q_i$ with $i = 1, 2$ and substituting $\beta_i = 1/k_B T_i$, Equation (3.42) may be re-written as

$$\frac{Q_1}{T_1} + \frac{Q_2}{T_2} \geq 0. \tag{3.44}$$

It is then straight forward to generalise this for $i = n$ and so:

$$\sum_i \frac{Q_i}{T_i} \geq 0. \tag{3.45}$$

Together with Equation (3.39), we then have derived that

$$\Delta S \geq 0 \tag{3.46}$$

for processes of the above kind.

The above equation requires further explanation. Why exactly is the von Neumann entropy of a quantum system non-decreasing? The reason lies in the subadditivity of the von Neumann entropy,

$$S(\rho_{12}) \leq S(\rho_1) + S(\rho_2) \tag{3.47}$$

and the fact that it is conserved under unitary evolutions. To demonstrate this, let us consider two initially uncorrelated systems[27] $\rho_{12} = (\rho_1 \otimes \rho_2)$ that are allowed to evolve according to some unitary U. The system then evolves into ρ'_{12}

$$\rho_{12} = (\rho_1 \otimes \rho_2) \rightarrow U(\rho_1 \otimes \rho_2)U^\dagger = \rho'_{12}. \tag{3.48}$$

As was just pointed out, unitary evolution is entropy-preserving and so

$$S(\rho_{12}) = S(\rho'_{12}). \tag{3.49}$$

[27]The fact that the systems are taken to be uncorrelated initially is essential for this, and resembles Boltzmann's famous *Stosszahlansatz*, in which he takes the velocities of the molecules of a gas to be uncorrelated in order to derive the H-theorem (Boltzmann, 1970; Ehrenfest and Ehrenfest, 1909; Brown et al., 2009).

Due to subadditivity it can then easily be seen that

$$S(\rho_{12}) = S(\rho_1) + S(\rho_2) \leq S(\rho_1') + S(\rho_2') \tag{3.50}$$

with equality in case of a vanishing interaction term in the joint system's Hamiltonian (including interactions due to the unitary), in which case the systems remain uncoupled. If the joint system is in a pure state, the von Neumann entropy vanishes completely.

The above equations show that after an uncorrelated quantum system is allowed to interact with another system and the two become entangled, the marginal von Neumann entropies of the two systems increase[28]. In the case of thermalisation, the system approaches the canonical state at the temperature β of the heat bath. Here, the same thing happens: the system is allowed to interact with different bath subsystems, and so the marginal state becomes closer and closer to the canonical state. This tracing out of the heat bath degrees of freedom after each interaction is a type of *coarse-graining* and is responsible for the irreversibility observed in quantum thermodynamics.

Example: Isothermal Process

I'd now like to discuss a particular example of processes, namely isothermal processes. It will be shown that here as well, the von Neumann entropy emerges as the quantum statistical generalisation of the thermodynamic entropy. Just like before, we consider a system with internal Hamiltonian H in contact with a single heat bath at temperature $\beta = 1/k_B T$. The system is taken to be in a canonical state and at

[28] As was already mentioned, the unitary evolution *also* means that there will be recurrence for finite systems.

equilibrium with the heat bath. Upon varying the Hamiltonian of the system, we perform an isothermal process and obtain for the average work change

$$W = \int_{t_0}^{t_1} Tr\left[\frac{\partial H}{\partial t}\frac{e^{-\beta H}}{Tr\left[e^{-\beta H}\right]}\right] dt \tag{3.51}$$

$$= \int_{t_0}^{t_1} \sum_n \frac{\partial E_n}{\partial t}\frac{e^{-\beta E_n}}{\sum_n e^{-\beta E_n}} dt \tag{3.52}$$

$$= -kT \ln\left[\frac{Z(t_1)}{Z(t_0)}\right], \tag{3.53}$$

where $Z(t) = Tr\left[e^{-H(t)/kT}\right]$ is the partition function. The average change in the system's energy is then $\Delta E = \langle H(t_1)\rangle_{\rho(t_1)} - \langle H(t_0)\rangle_{\rho(t_0)}$ and the system transfers heat $Q = W - \Delta E$ to the heat bath. For a canonical thermal system, that is for $E_n = -kT \ln Z - kT \ln[p_n]$, where p_n are the system's eigenvalues, this then results in

$$Q = kT\left(\sum_n p_n(t_1)\ln[p_n(t_1)] - \sum_n p_n(t_0)\ln[p_n(t_0)]\right). \tag{3.54}$$

$$= T\left[S(\rho(t_1)) - S(\rho(t_0))\right] \tag{3.55}$$

where $S(\rho) = -kTr[\rho \ln \rho]$ is the familiar von Neumann entropy.

This example shows that it is possible to derive the expression for the von Neumann entropy, simply by considering manipulations of a joint system's Hamiltonian and by making use of the properties of the canonical state.

In particular, there was thus far no need to equip the above formalism with an interpretation, and so the derivation of thermodynamics from quantum mechan-

ics demonstrably does not require considerations about ignorance, or information. Hemmo and Shenker's claim that von Neumann's argument for his entropy "is unique because it is the only justification offered so far for the idea that $S(\rho)$ is entropy" (Shenker, 1999, p.35) is therefore not correct. As Maroney (2007) points out repeatably, once the canonical state has been accepted as representing thermal states, the equations of quantum thermodynamics follow in a fairly straight forward manner. And this motivation for considering the canonical state as the thermal state has been given by reflecting on the properties of heat baths, hence by using phenomenological reasoning.

There remains one potential issue with this approach, however: it only refers to *average* quantities. This will be addressed in the next section.

3.2.5 Interpretations of Heat, Work and Entropy

The above derived quantum thermodynamic laws are all given in terms of expectation values. How we understand heat, work and entropy in quantum mechanics therefore hinges fundamentally on our understanding of these expectation values. In this section I will first briefly discuss the use of expectation values in a thermodynamic setting before moving on to an analysis of how one might understand these expectation values in the given quantum mechanical context.

A well known problem with expectation values is that, despite their name, for a single experiment those expectation values may not at all be what one may expect the outcome to be. Expectation values lack some of the important information that the probability distribution itself is able to provide us with[29]. For example,

[29] Although once $\langle f(x) \rangle$ is known for all functions f, this is equivalent to knowing the probability distribution over x.

if we consider an infinite (or sufficiently large) series of successive die throws, the expectation value not only fails to establish the fact that each number is equally likely (after all, this is what the probability distribution is for), but it furthermore calculates to be 3.5, which does not even resemble a physical value a standard die could take. Equation (3.22) alone therefore will neither be able to tell us how much work one can *actually* extract from a given individual quantum system, nor whether the amount of extracted work will be *close* to the calculated mean. It will rather tell us how much work we can extract *on average*. This fact alone might *prima facie* not be too troubling: the previous sections discussed at several points that the second is best understood in what I called a probabilistic version, a version that only prohibits the reliable conversion of heat into work. This notion is made precise (and tightened) by the use of expectation values, as was seen above[30]. One may reject such version of the second law, but in that case thermodynamics would lose much of its power and scope.

The conceptual problems actually lure in the details of the averages: an average about *what*? Is it an average over the subsystems that are part of a larger system (but then what happens in the case of a *single* system)? Or an average over a large number of repeated cycles? Is the mean a property of the system itself or not? In very large (by which I mean thermodynamic-limit-large) systems, fluctuations are absent and so the average values of heat, work, entropy and energy will basically coincide with the actual values. In the case of smaller systems, however, fluctuations become of importance and with them the question of how to interpret those averages.

Chapter 1 explained some of the common criticism that is associated with a Gibbsian understanding of statistical mechanics. The main case against the Gibbs entropy was

[30]Maroney (2009a) points out that these two versions (reliable vs expected) are non-equivalent, because in the first version, there exists the possibility that the demon could violate the second law on any *finite* time-scale.

shown to be the claim that the probabilities that appear in the formulation of the entropy are most naturally understood to be measures of ignorance. This was due to the intuition (and arguably Jaynes' 1957 influential account which shaped today's standard conception in physics) that due to the underlying deterministic dynamics, the probabilities quantify our ignorance over the system's underlying microstates, instead of being 'out in the world'. But now we are in a quantum mechanical setting! Heat, work and entropy are now functions of the quantum state ρ, and not functionals of a probability distribution over classical phase state. Instead of probabilities that are added on top of an underlying deterministic theory, quantum mechanics *itself* is probabilistic. And so the probabilities arise from the Born rule, and not from considerations about ignorance.

Having said that, there *are* interpretations that maintain that the quantum state is best understood as quantifying an agent's ignorance, Quantum Bayesianism (QBism) is one of the recently developed accounts (Caves et al., 2002, 2007; Fuchs and Peres, 2007; Fuchs et al., 2013). Within a QBist framework, the thermodynamic entities above will naturally be understood in terms of an agent's epistemic capacities, as a trivial consequence from the much more severe claim that the quantum state *itself* merely represents an agent's epistemic state, a collection of subjective degrees of belief (Fuchs, 2002). While the QBist approach has been criticised by philosophers for independent reasons (Timpson, 2008; Brown, 2017), it certainly gives no relief to those who are worried about entropy being subjective. Quite the contrary. QBism is concerned with predictions, but as Timpson (2008) points out, cannot offer us any explanations of physical phenomena. By extension the QBist will not be able to provide us with explanations of thermodynamic phenomena. Entropy then is nothing deeper than a predictive tool for the impossibility of extracting work. This itself is of course in no conflict with thermodynamics itself, it just commits to one particular

instrumental reading of it and delivers little explanation.

From the above it becomes clear that our understanding of quantum mechanics directly influences our understanding of the meaning of the thermodynamic quantities derived in this chapter. Classical statistical mechanics is a theory about probability distributions, the above account of thermodynamics in the quantum regime on the other hand arose from considerations about mixed quantum states. The origin of the probabilities that occur in the quantum mechanical description of heat, work and entropy, are therefore determined by our understanding of probabilities in quantum mechanics. In Everettian quantum mechanics, for example, the quantum state evolves deterministically and the probabilities are emergent properties of the decoherent branching structure of the universe (Wallace, 2012). How to interpret these probabilities remains somewhat controversial, from suggestions reaching from interpreting them as general uncertainties (Saunders, 1998) or post-measurement-outcome uncertainties[31] (Vaidman, 2012) (see also (Vaidman, 2016) and references therein). It should be noted however, that these probabilities arise from the underlying theory itself and are not only, as Wallace (2013a) points out, an "epiphenomenal gloss on underlying physics" (p.17). This means that even though entropy in turn may be associated with uncertainty, this uncertainty is of a very different nature as the ignorance encountered in the classical case. It is a feature of quantum theory itself that tracks an objective property of the system in a decohering context.

For other accounts, the fundamental laws themselves are stochastic, as for example in the case of GRW (Ghirardi et al., 1986) dynamical collapse theory[32]. Here the probabilities become part of the dynamical laws of the system, and are therefore objective and non-epistemic. Frigg and Hoefer (2007) for example identify the

[31]The quantitative argument for why the probabilities are in accordance with the Born rule is usually made by decision-theoretic considerations. (Deutsch, 1999; Wallace, 2003; Greaves, 2007)

[32]Standard quantum mechanics needs to be augmented in this case.

two possible interpretations of probabilities in a GRW context to be either single-case propensities or Humean objective chances. Finally, there are also deterministic hidden variable theories[33], such as Bohmian mechanics (Bohm, 1952; Dürr et al., 2012; Goldstein, 2017), where the wave function is only a partial description of the quantum system. A 'guiding equation' is thereby added to the quantum formalism, that describes how the Bohmian corpuscle's position and momentum evolve. The evolution of the corpuscle is fully deterministic, and so it seems to me that understanding these probabilities in objective terms is similarly difficult as in the classical case.

To summarise the above: how we understand the thermodynamic notions of heat, work and entropy in the quantum context depend entirely on how we interpret the mixed state. Wallace (2013a) also discusses the case of quantum *statistical mechanics*, in which case the question arises whether and how one should draw a one-to-one comparison between quantum statistical mechanics and statistical mechanics by identifying Hilbert space with phase space, pure states with microstates and equipping Hilbert space with an additional probability distribution. I agree with the conclusion of his analysis that the need for adding (and interpreting) such an additional probability measure over quantum states (in need of interpretations themselves) is unnecessary, and that in fact quantum theory itself "will do just fine" (p.19).

3.2.6 Quantum Resource Theories

This last section will deal with a rapidly expanding area of contemporary physics: quantum resource theories. Resource theories have sprung up like mushrooms in the last few years: there now exists a resource theory of entanglement, a resource theory

[33]I have not mentioned stochastic hidden variable theories, but there similar arguments as in the GRW case can be made.

of quantum reference frames, a resource theory of non-Gaussianity, a resource theory of asymmetric quantum states, and many more (Weedbrook et al., 2012; Horodecki et al., 2009; Bartlett et al., 2007). The focus here will lie on the resource theory of non-thermal quantum states, as it is thought to have "advanced our understanding of fundamental physical principles, such as the second law of thermodynamics" (del Rio et al., 2015, p.1). In particular it is claimed that thermodynamics can be derived from such resource-theoretic considerations, given that "the free energy of thermodynamics emerges naturally from the resource theory of energy-preserving transformations" (Brandão et al., 2013, p.1).

Before going into further details, it is useful to briefly recall what resource theories are and how they function. In short, all resource theories are made of three basic ingredients (Brandão and Gour, 2015):

(i) The set $\mathcal{S}$ of 'free states' that are freely and abundantly available to the experimenter,

(ii) a restricted set $\mathcal{C}$ of available quantum operations which is closed under $\mathcal{S}$ and

(iii) a set $\mathcal{R}$ of 'resources', which consist of states that cannot be created by $\mathcal{S}$ and $\mathcal{C}$ alone.

As a consequence, any state that is not in $\mathcal{S}$ can only be accessed by making use of resource states. The various resource theories then differ in what they consider to be the free states, the allowed operations and the resources. In the resource theory of entanglement, for example, the free states are the separable states, the resources are the entangled states and the (maximal) set of allowed operations are

the LOCC operations[34]. In the resource theory of thermodynamics, on the other hand, the free states are the canonical states at temperature T, the resources are all non-thermal states and the set of allowed operations are given by the completely positive trace-preserving maps $\mathcal{E} : \mathcal{L}(\mathcal{H}) \to \mathcal{L}(\mathcal{H})$ for which

$$\mathcal{E}(\rho) = Tr_B \left(U_{AB} \left(\rho_A \otimes \rho_{B,\beta} \right) U_{AB} \right), \tag{3.56}$$

where $\rho_{B,\beta}$ is a thermal state at temperature $\beta = 1/k_B T$ and U_{AB} is an arbitrary unitary operation that commutes with the total Hamiltonian (Brandão et al., 2013). Tracing out degrees of freedom therefore is also counted to the set of allowed operations.

Within each resource theory it is possible to derive what is called a resource-monotone R. This is a function that obeys a partial order which in turn determines which resources can be transformed into each other via the application of the set of allowed operations. Formally, for two resources A and A', one writes $A \geq A'$ if the set of free operations allows for the transformation of A into A'. In the resource theory of thermodynamics, the monotones are given by the relative entropy:

$$D(\rho_A || \rho_{B,\beta}) = Tr[\rho_A \ln \rho_A - \rho_A \ln \rho_{B,\beta}]. \tag{3.57}$$

This entropy is non-decreasing and it is necessarily the case that

$$D(\mathcal{C}(\rho_A) || \rho_{B,\beta}) \leq D(\rho_A || \rho_{B,\beta}). \tag{3.58}$$

[34]LOCC stands for 'local operations and classical communication'. It classifies a method in which the results of any locally performed operations is communicated classically to another agent, who in turns may then perform another local operation.

The relative entropy can be re-written in terms of the free energy (Brandão et al., 2013) $F_\beta(\rho) = \langle H \rangle_\rho - 1/\beta S(\rho)$, It characterises the average amount of work one can extract from a system. $S(\rho)$ is the von Neumann entropy, as above. It then becomes:

$$D(\rho_A || \rho_{B,\beta}) = \beta F_\beta(\rho_A) - \beta F_\beta(\rho_{B,\beta}), \tag{3.59}$$

And so it seems indeed as if the free energy emerges naturally from the resource theory of quantum thermodynamics. From the division of states into the two sets of thermal and non-thermal states, and from the restriction of unitary operations to those that are energy-conserving, it is therefore possible to derive the second law.

In a way, the resource theory of thermodynamics seems to fit well with what was called an operational understanding of thermodynamics in the first section of this article. If thermodynamics is about which processes a system can undergo given a set of fixed operations, then this is exactly what the resource theory is giving us. But one may justifiably wonder whether resource theories aren't too broad for the purpose of gaining insight into something as profound as the second law of thermodynamics. After all the resource monotones are generic features of the theory, and the theory is applicable to a wide variety of fields, only a fraction of which has anything to do with thermodynamics.

The key is that the heavy lifting has already been done before the resource theoretic machinery grasps. Yes, thermodynamics emerges from the formalism, but only because we put thermodynamics *in* before. The identification of thermal states with the canonically distributed states, the existence of large reservoirs at different temperatures that contain these thermal states and the choice of energy-conserving unitaries as the set of allowed transformations — it has been shown above that these ingredients are already enough to derive the full set of thermodynamic equations

from quantum mechanics. The resource theory of thermodynamics therefore captures the operational spirit of thermodynamics but it does not suffice to explain thermodynamics. This however does not mean that there are no insights to be gained from resource theories, quite on the contrary. With their lean structure and wide applicability, they systematically show how monotones emerge from a limited set of ingredients. In particular they show that these monotones are generic features of those ingredients. Understanding this relation between monotones and ingredients might therefore explain thermodynamic-like behaviour in other areas of science.

Single Shot Thermodynamics

The thermodynamic behaviour of quantum systems above was derived in terms of expectation values. Sometimes, however, it is not averages that we are interested in, but the *actual* amount of work one can extract from a quantum system in a single experiment. This is where single shot thermodynamics enters the picture (Dahlsten et al., 2011; Horodecki and Oppenheim, 2013; Skrzypczyk et al., 2014; Brandão et al., 2015). The arguments in the literature are usually made in terms of the free energy. In the phenomenological case, a necessary condition for the extraction of work is a decrease in the system's *free energy*:

$$F(\rho) = \langle H \rangle_\rho - TS(\rho), \tag{3.60}$$

The canonical distribution minimises the free energy, and so for a system in contact with a heatbath and at equilibrium, this reduces to

$$F(\rho_\beta) = -kT \ln Z. \tag{3.61}$$

The free energy restricts the sets of admissable transformations from ρ to ρ' when a system is in contact with a heat bath, i.e. at constant temperature. The fact that the free energy is is non-decreasing can be regarded as a version of the second law (Sagawa and Ueda, 2008). In particular, the change in free energy provides us with a limit for the maximal amount of work that can be extracted from the system:

$$W \geq \Delta F(\rho) \tag{3.62}$$

This inequality holds for systems of any size and it does not need to be assumed that the system ends up in a thermal state at the end of the transformation—in fact, it could end up in any non-equilibrium state and equation (3.62) would still hold. Horodecki and Oppenheim (2013) recently showed that equation (3.62) is a bad bound for *individual* quantum systems. In practice, much less work can be extracted. As a result, Brandão et al. (2015) have developed a *whole family* of second laws that restrict the range of allowed processes (allowing however for the possibility of using a so-called *catalytic* state that can be used during the process but has to be brought back to its initial state by the end of it). There was much excitement about these findings in the research community — in particular because they seemed to confirm that our image of the relation between work and free energy or entropy on the phenomenological level was deceitful and that on the fundamental level the free energy is not suitable as a measure of extractable work. "[...] the free energy [equation (3.60)] is only valid in the thermodynamic limit" (Horodecki and Oppenheim, 2013, p.2). The analogy between free energy difference and extractable work therefore can only be recovered when considering operations acting on many copies of the system at the same time.

The above results may be considered to be more of practical than of conceptual

relevance: if I am interested in the amount of work that can deterministically be extracted in a *single* experiment, then naturally this will amount to much less than what is given by the expectation values. It is also unsurprising that such an approach will lead to a whole family of second laws, as the work extractable from a system by a single operation is highly context dependent. This may be of great practical importance, but calling these newly found bounds for single-shot work extraction 'second laws' seems somewhat odd, as it suggest that there is some deep underlying regularity that governs these processes. But this is not the case: the 'second laws' derived by Horodecki and co., as useful as they may be for all practical purposes, given their context-dependent applicability lack the universal aspect that we encounter for example in the quantum mechanical second law introduced earlier in this section, which is always true (assuming quantum mechanics is true).

This brings us to the question of what Horodecki and co. take the old laws of thermodynamics to be about. As can be seen from the above quote, for them, the laws are only valid in the thermodynamic limit when fluctuations become negligible. This is very similar to Maxwell's statistical view of the second law[35]. It results from the fact that they are once again concerned with practical matters: for the actual work extraction to become identical to the expectation value, one would require an infinite amount of systems[36]. Their statement that the laws of thermodynamics are only valid in the thermodynamic limit may then be traced back to either the expectation that thermodynamics ought to give us exact predictions about how much work one can extract in a single experiment. Or it may be traced back to a particular interpretation of the expectation values in frequentist terms, which, as

[35]Minus the fact that Maxwell thought arbitrary violations could be possible.

[36]Although here the problem discussed in the context of frequentism enters again: there is no guarantee that the average work extracted will agree with the expectation value, it is just very likely that they do.

was shown above is certainly not necessary in the quantum context. In a way, this view resembles that of Hemmo and Shenker, who as was shown above, take the von Neumann entropy to only be identifiable with the thermodynamic entropy in the infinite particle limit.

3.2.7 Conclusion

It has been shown that the laws of thermodynamics emerge naturally from the quantum formalism once the distinction between two types of energy transfer (heat and work) is made, and once the canonical states are identified with the thermal states. The latter can be justified by considering passive states as the only states that are feasible for modelling an ideal heat bath. The fact that the quantum mechanical laws of thermodynamics can be arrived at by applying either dynamical insights or by using resource theory, is furthermore remarkable. It suggests that at least the *structure* of the laws of thermodynamics may not be all that special after all. I also very briefly discussed some aspects of single-shot thermodynamics, as it demonstrates further how the laws of thermodynamics are interpreted differently across the areas.

In the next part, I will consider what happens when quantum thermodynamics, and in particular Landauer's principle, are applied to the foundations of quantum mechanics.

3.3 Entropy and the Foundations of Quantum Mechanics

In this last part of the chapter, I will explore the idea whether thermodynamics can help us gain a deeper understanding about the foundations of quantum mechanics. For this, I will discuss an article published by Cabello et al. (2016), who maintain that thermodynamics allows us to distinguish between two broad classes of quantum interpretations[37], one of which is supposedly shown to be untenable. This section has been published as a co-authored paper (Prunkl and Timpson, 2018).

3.3.1 Introduction

For nearly a century, physicists and philosophers alike have puzzled over how to interpret quantum theory, unable to decide unambiguously between a variety of more or less promising candidates. In a recent publication, Cabello et al. (2016) put forward an argument which seeks to demonstrate the existence of a real, empirical difference between various interpretations of quantum mechanics. Moreover, the authors assert that it is in principle possible to measure this difference experimentally. Their argument is based on methods derived from computational mechanics—a growing field that is concerned with the simulation and prediction of stochastic processes. Interestingly, when applied to certain physical processes, computational mechanics is able to provide us with thermodynamical limitations on these processes (Wiesner et al., 2012; Garner et al., 2015). Cabello et al.'s argument is a concrete, foundationally motivated, application of computational mechanics which suggests

[37] I take Bohmian Mechanics and GRW to be interpretations of quantum mechanics as well, even though strictly speaking, they are theories in their own rights.

that there is a thermodynamical cost to bear for a subset of quantum interpretations: perhaps a pathological one.

The link between thermodynamics and computational mechanics can be understood as follows: depending on the complexity, i.e., randomness, of a pattern that is to be simulated, greater or fewer resources are needed in order either to create the pattern or to predict its future, given observations of past data sequences. By 'pattern' we simply mean a time-series of data points, for example measurement outputs. One can take the computational system we are interested in simulating to be a black box, with the only accessible empirical data being its input and output variables. It can then be proven that there exists a machine, called an ϵ-machine, which is predictively optimal and uses the minimum resources, while simulating the input-output behaviour of the target system (Crutchfield and Young, 1989). For some thermodynamic systems this method shows up the limitations for work extraction via physical processes. Given a resource-theoretic understanding of thermodynamics (i.e., an understanding which conceives thermodynamics primarily to be a theory about what tasks one can perform when furnished with certain resources[38]), one might say that computational mechanics can be considered a useful tool for the task of understanding and enhancing the foundations of thermal physics.

Cabello et al. begin by dividing the set of quantum interpretations into two subsets, what they term *Type I* and *Type II* interpretations. They then argue that Type I interpretations are associated with a thermodynamical cost, rendering such interpretations highly problematic: either one of the (very plausible) three assumptions must be given up, or there exists a surprising heat generation which could be ruled in or

[38]Modern accounts include (Horodecki and Oppenheim, 2013; Wallace, 2014; Brandão et al., 2015; Gour et al., 2015), however, the underlying idea that thermodynamics is to be understood relative to an agent and her means goes back to Maxwell (1871) (c.f. Myrvold (2011)) and was later famously promoted by Jaynes (1965).

out experimentally (and one would be surprised indeed if such heat generation were in fact to be found). If correct, this result would seem an outstanding breakthrough whose far reaching consequences might not only force us to abandon some of the most popular interpretations of quantum mechanics (Type I interpretations include such favourites as de Broglie–Bohm theory, Everett, and dynamical collapse theories such as GRW, for example) but would shake the foundations of our understanding of the relationship between scientific theories and the underlying ontic structure of the world.

Our prime concern in this paper is to assess Cabello et al.'s argument and the tenability of their conclusions. But we also have their example in mind as a test-case for the application of computational mechanics in pursuit of dividends in foundations of physics.

We will begin with a brief outline of stochastic input-output processes before presenting the argument of Cabello et al. (2016), which applies this mathematical machinery to quantum systems. We will then analyse why Cabello et al.'s argument about the thermodynamical costs of some quantum interpretations fails, including offering a straightforward counterexample. We will show that the adumbrated heat cost is in fact not associated with the quantum system itself at all—it is not of quantum origin—and thus controversies over quantum interpretations are not germane to it, nor it to them. Rather, the heat cost arises with the external experimental setup stipulated by Cabello et al.

We begin by introducing the most important aspects of stochastic input-output processes, as they form the backbone of the argument. More detailed discussions may be found in (Barnett and Crutchfield, 2015; Crutchfield and Young, 1989).

3.3.2 Computational Mechanics: Input-Output Processes

The goal behind modelling systems' behaviour by input-output processes is to find the minimal structural requirements that produce a particular statistical pattern. To do so, one works backwards from the statistics of experimental outputs to then find the minimal amount of resources needed in order to simulate output strings that are statistically indistinguishable from the actual experimental result.

More formally: A stochastic process $\overleftrightarrow{Y}$ is described as a bi-infinite one-dimensional chain $..., Y_{-1}, Y_0, Y_1, ...$ of discrete random variables $\{Y_t\}$ with values $\{y_t\}$, where t is a discrete time parameter and the direction of the arrow above the random variable indicates whether the chain extends to the past (left arrow), the future (right arrow) or to past *and* future (left-right arrow) infinity. The $\{y_t\}$ are the particular values the random variable takes at time t and in our case we can think of them as the output values of an experiment performed on the system. For example, for a spin-measurement on a qubit—the kind of case with which Cabello et al. will be concerned—the outcome-types could be "up" and "down" for example, taken from the output alphabet $\mathcal{Y} = \{$"up","down"$\}$. If not only the output but also the input is stochastic (in our case, this will correspond to a choice of spin-measurement basis, which will be taken to be random) the effect of the input random variable on the future statistics needs to be taken into account as well. Such a process must then be modelled by a so-called stochastic *input-output process*, $\overleftrightarrow{Y}|\overleftrightarrow{X}$, with input values $\{x_t\}$ from an alphabet $\mathcal{X}$. The whole input-output process may then be described as a collection of stochastic processes $\overleftrightarrow{Y}|\overleftrightarrow{X} \equiv \{\overleftrightarrow{Y}|\overleftrightarrow{x}\}_{\overleftrightarrow{x}\in\overleftrightarrow{\mathcal{X}}}$, where we take each process $\overleftrightarrow{Y}|\overleftrightarrow{x}$ to correspond to all possible output sequences $\overleftrightarrow{Y}$ that could arise from one particular input sequence $\overleftrightarrow{x}$, drawn from the set of all possible input sequences $\overleftrightarrow{\mathcal{X}}$.

The *probability distribution*[39] over the set of all possible output sequences, given a particular input sequence, is then given by what is called the *channel's distribution*:

$$\mathbf{P}(\overleftrightarrow{Y}|\overleftrightarrow{x}) = \{\mathbf{P}(\overleftrightarrow{Y} \in \sigma|\overleftrightarrow{X} = \overleftrightarrow{x})\}_{\sigma \subseteq \overleftrightarrow{\mathcal{Y}}, \overleftrightarrow{x} \in \overleftrightarrow{\mathcal{X}}} \tag{3.63}$$

The idea is now to divide the input-output sequences into *pasts* and *future* and furthermore to divide the various input-output pasts into sets that yield the same distribution over input-output futures. Two input-output pasts $\overleftarrow{z} = (\overleftarrow{x}, \overleftarrow{y})$ and $\overleftarrow{z}'$ that yield the same future input-output conditional probabilities $P(\overrightarrow{Y}|\overrightarrow{X}, \overleftarrow{Z} = \overleftarrow{z}) = P(\overrightarrow{Y}|\overrightarrow{X}, \overleftarrow{Z} = \overleftarrow{z}')$ are then said to belong to the same *causal state* s. Denote the set of causal states $\mathcal{S}$. The ϵ-map is then introduced as the mapping $\epsilon : \overleftarrow{\mathcal{Z}} \to \mathcal{S}$ from any input-output past onto its corresponding causal state (Barnett and Crutchfield, 2015). This map also induces a probability distribution over the causal states, which since the process is stationary and ϵ is time-independent, is called the process' *stationary distribution*. The causal states contain all the relevant information for optimally predicting the future output statistics of the system and contain as much information as any of its input-output pasts. For simplicity we take the input sequences to be uniformly distributed. The minimal amount of information needed to be stored in order to predict future outputs optimally is then given by the Shannon information $H(\mathcal{S})$ and is called the *statistical complexity*. This also quantifies the amount of resources needed in order to model the system's future behaviour.

3.3.3 Foundations: Division of Interpretations into two Groups

Cabello et al. (2016) seek to use the above machinery in combination with a few

[39] We consider only stationary probabilities, which means that the probabilities are time translation invariant.

plausible assumptions to raise difficulties for a group of well-known quantum interpretations. Their approach is to divide the set of quantum interpretations into two broad classes, based on their respective takes on quantum probabilities: Type I interpretations are interpretations that regard probabilities as determined by "intrinsic properties of the system" (Cabello et al., 2016, p.1). These properties typically change post-measurement, depending on the choice of measurement performed on the quantum system. Examples which they mention of Type I interpretations include de Broglie-Bohm theory (Bohm, 1952; Goldstein, 2017), many worlds interpretations, e.g. (Everett, 1957; Wallace, 2012), Ballentine's statistical interpretation (Ballentine, 1970), modal interpretations (Lombardi and Dieks, 2016) and consistent histories, as well as GRW dynamical collapse theories (Ghirardi et al., 1986) and Spekkens' toy model (Spekkens, 2007).

Type II interpretations, on the other hand, comprise those interpretations which treat probabilities as "relational properties between an observer and the system" (Cabello et al., 2016, p.1). According to such interpretations, the quantum state corresponds to the "experiences an observer has of the observed system"(Cabello et al., 2016, p.1). To the class of Type II interpretations, the authors assign, amongst others, the Copenhagen interpretation, Wheeler's view (Wheeler and Zurek, 2014), relational interpretations (Rovelli, 1996) and QBist interpretations (Fuchs and Peres, 2007).

We note to begin that there are some significant problems associated with such a division of interpretations. First, some of the interpretations listed as Type I do not seem unambiguously to belong in either camp: the Everett interpretation, for instance, may well be taken to imply that probabilities ought to be regarded relationally as opposed to being intrinsic properties of a quantum system. After all,

the Everettian could argue, the 'bare quantum formalism' is fully deterministic at the fundamental level. The probabilities, by contrast, occur at an emergent level and justifications of the Born rule are typically given via decision-theoretic arguments (Wallace, 2012), very strongly suggesting that probabilities in Everett, if they need an interpretation at all, are best interpreted as concerning relational experiences of an agent within a given branch.

The Type I/ Type II distinction also seems problematic with regard to Ψ-epistemic interpretations, in which the quantum state is taken to represent information about how a system is[40]. For instance, the probabilities occurring in Spekkens' toy model—which allegedly belongs to the Type I camp—are purely epistemic and result from an incomplete knowledge about the underlying ontic state[41] (Spekkens, 2007), as opposed to being intrinsic properties of the system: they merely occur on the level of the observer whereas the underlying ontic evolution could be, for all we know, a deterministic one. It thereby seems somewhat misplaced to regard these probabilities as "observer-independent" (Cabello et al., 2016, p.1), in the sense of being part of the furniture of the world, a notion more commonly found in the context of propensity interpretations of probabilities (Popper, 2002), but certainly not in an epistemic context.

We may try to make clearer sense of the Type I/ Type II distinction by phrasing it differently in order to capture what Cabello et al. seem to have in mind: namely a distinction between interpretations that either endorse or reject the notion of an underlying ontic physical state. What the interpretations listed under Type II then have in common is that they all roughly follow along Niels Bohr's line: There is no quantum world. At least, there is not a free-standing mind- and agent- independent

[40]For details on the Ψ-ontic/Ψ-epistemic distinction, see Harrigan and Spekkens (2010).

[41]In analogy with Jaynes' view of probabilities in classical statistical mechanics, where the probability distribution over the underlying ontic state equally represents ignorance (Jaynes, 1965)

one. Type II interpretations thereby operate on a completely different level from Type I interpretations, the latter of which consider quantum mechanics as ultimately describing a free-standing mind-independent objective reality. Drawing a distinction between these fundamentally different conceptions of scientific theory on the basis of their conception of probabilities, however, is misleading and unsatisfying. The probabilities are mere epiphytes on a deeper, more pressing issue about the nature of science.

Consequently, if Cabello et al. are indeed correct and Type I interpretations could be ruled-out experimentally, then their result would have profound consequences: it undermines one of our core traditional conceptions of scientific practice, namely that scientific theories tell us what the world independent of us is like.

3.3.4 Three Assumptions

As Cabello et al. explain, their argument concerning Type I interpretations rests on three assumptions, which we can state as follows:

(i) Which measurement is performed on a system is decided randomly, and in particular independently of the state of the system.

(ii) A (finite dimensional) quantum system has a limited memory.

(iii) Landauer's principle is valid.

They go on to suggest that for Type I interpretations, it follows from (i) that a system's *intrinsic properties* typically ought to *change*, depending on which measurement is performed on the system. When the authors speak of 'limited memory', (ii), they seem to have in mind something like the following: when a system is being measured

in some basis, it typically generates a new value of its intrinsic property or set of intrinsic properties, which will determine the quantum probabilities of the future measurements. Measurement in a different basis will force the system to change these values in order to comply with the correct quantum probabilities, thereby deleting—or perhaps rather overwriting—the previous values. Due to its limited memory capacity, the system cannot possess all the intrinsic values that will determine the probabilities of all possible measurement series. We may assume that by 'intrinsic properties' Cabello et al. have in mind properties that determine the probabilities of future measurement outcomes as opposed to properties specifying pre-determined definite outcomes, the reason being that this latter interpretation of "intrinsic properties" would force us to adapt some sort of hidden variable interpretation, which, by definition is only a subgroup of the Type I class. We therefore take it that the former meaning of 'intrinsic properties' is intended.

Preliminary assumption (iii) is the validity of Landauer's principle. As it is often stated, Landauer's principle asserts that the erasure of information in a system's information bearing degrees of freedom is accompanied by an increase of entropy in the non-information bearing degrees of freedom (Landauer, 1961). This increase of entropy will lead to the dissipation of $kT \ln 2$ of heat per deleted bit, where T is taken by Cabello et. al. to be the temperature of the system (though more usually it is taken to be the temperature of the environment) and k is, of course, the Boltzmann constant.

3.3.5 Heat Dissipation between Successive Measurements

We are now invited to consider a single qubit, on which an observer performs successive projective measurements in one of the two Pauli-bases, σ_x or σ_z, chosen

at random with equal probability (assumption (i)). Since the state of the qubit changes after successive measurements in non-orthogonal bases, within a Type I framework the internal properties of the system must change too. Given that the system has limited memory (assumption (ii)), it needs to generate new values that determine its future behaviour and store them in its memory. To do so, the system will need—as they put it—to 'erase information'. Applying Landauer's principle (assumption (iii)), Cabello et al. (2016) finally conclude that during this erasure process, heat is dissipated into the environment.

To quantify the amount of information that needs to be erased, they model the system as a computational machine, a black box that generates output strings on the basis of some input and its internal memory. The optimal and minimal machine that is able to produce output strings that are statistically indistinguishable from the actual experimental outcomes, will be, as we mentioned earlier, an ϵ-machine (Crutchfield and Young, 1989). It maximises the mutual information between input-output past and output future, thus simulating the statistical process, and with minimal resources.

Applying the computational mechanical machinery to the qubit in question, we identify the input random variable X_t with the choice of measurement basis, randomly selected from the alphabet $\mathcal{X} = \{\sigma_x, \sigma_z\}$. The output variable Y_t represents the measurement results and can take values ± 1. The causal state after the measurement is simply taken to be the respective quantum state, $s_o = |0\rangle$, $s_1 = |1\rangle$, $s_+ = |+\rangle$ or $s_- = |-\rangle$ (with, note, no specific view here taken on the ontology or otherwise, of the quantum state).

The machine has a probability 1/2 of changing its causal state after a measurement. Hence, half of the time, it must update its internal properties, and Cabello et al

maintain that this requires the erasure of information. They say *"The average information that must be erased per measurement is the information contained in the causal state previous to the measurement, S_{t-1}, that is not contained in the causal state after the measurement, S_t."* (Cabello et al., 2016, p.2). It should be noted that this formulation is perhaps somewhat misleading, as it suggests that *a particular* causal state itself carries a certain amount of information. In fact, it is not *the* causal state to which we assign an entropy, but it is instead the probability distribution over causal states which is associated with an entropy and thereby with a Shannon information. Such an average over causal states however is not a causal state itself.

In any case, the amount of information that needs to be erased is equal to the *conditional entropy* of

$$I_{erased} = H(S_{t-1}|X_t, Y_t, S_t). \tag{3.64}$$

(See Appendix.)

In the given experiment, the probability distribution over the causal states is uniform, which allows us to only consider a particular causal state s_0, brought about by a particular measurement σ_z in order to determine I_{erased}. Together with the fact that $H(Y_t|S_t) = 0$, the average erased information for the case of a successively measured qubit is then calculated to be

$$I_{erased} = H(S_{t-1}|\sigma_z, s_0) = -\sum_{s_j \in \mathcal{S}} P(S_{t-1} = s_j|\sigma_z, s_0) \log P(S_{t-1} = s_j|\sigma_z, s_0). \tag{3.65}$$

The three possible causal states at time $t-1$ are s_0, s_+ and s_-, with transition

probabilities 1/2, 1/4 and 1/4.[42] The conditional entropy then turns out to be $I_{erased} = -\frac{1}{2}\log\frac{1}{2} - 2\cdot\frac{1}{4}\log\frac{1}{4} = \frac{3}{2}$ bits.

Cabello et al. (2016) thereby conclude that once we accept assumptions (i)–(iii), it follows that the system on average must dissipate $\frac{3}{2}kT\ln 2$ units of heat per measurement, if understood as belonging to Type I.[43] In principle, the above experiment could be implemented in a lab. From our previous observations of measurements on quantum systems, it is however safe to say that we would be very surprised indeed to observe any such heat dissipation. Cabello et al. thereby suggest that Type I interpretations are unlikely to be representative of the world, at least if their plausible assumptions (i)–(iii) hold. The above argument supposedly does not apply to Type II interpretations, however, as for these interpretations "measurement outcomes are created randomly when the observables are measured, without any need to overwrite information in the system and therefore without the system dissipating heat due to Landauer's principle" (Cabello et al., 2016, p.3).

In the conclusion of their paper Cabello et al. do canvass the possibility that one or more of assumptions (i)–(iii) might be thought to fail instead of Type I interpretations being lumbered with an excess heat cost, and in particular they judge that the de Broglie–Bohm theory and the Everett interpretation should not have the excess heat quantity attached to them, as both interpretations violate (ii)—the finite memory assumption. As they see it, in the de Broglie–Bohm case this is because the ontology includes a continuous field (the unitarily evolving wave function), and in the Everett case, because one has a splitting into a plurality of worlds. But it is at best obscure that either de Broglie–Bohm or Everett should be thought to

[42] The authors write at time t, but must have intended $t-1$.

[43] The authors furthermore generalise their result to N-outcome measurements with an associated heat generation which scales linearly with N. Thus sufficient measurements on a single system would produce as much heat as you like. This does not sound promising for Type I interpretations.

violate the finite memory assumption. In the de Broglie–Bohm case, from the fact that the quantum state is taken to be real[44] it does not follow that a system has infinite memory capacity: nearly all the state is irrelevant to the time evolution of the definite physical quantities most of the time anyway due to decoherence, whilst a particle's motion is only guided by the value of the wavefunction assigned to the region immediately surrounding it in any case[45]. And one might note that realist collapse theories such as GRW *also* have a continuous field in them: what difference could it make to judgements of memory capacity whether the continuous field (sometimes) jumps around stochastically (GRW) or whether it instead evolves determinisically but most of it being irrelevant to the evolution of a particle (de Broglie–Bohm)? With regard to Everett: even if there is branching, each world would be one in which the finite memory condition held[46], so each world would be one in which the excess heat cost obtained. If anything, this would look worse, rather than better, for Everett. In fact, Cabello et al. are quite right that there is no excess heat cost which arises for de Broglie–Bohm or for Everett, but this is not for the reason they allege (the possibility of infinite memory capacity). Rather it is, as we shall see, an instance of a general proposition: there is no excess heat cost for any Type I interpretation over a Type II interpretation.

Let us now analyse how Cabello et al. arrive at their surprising results. We will begin with a brief recapitulation of Landauer's principle before delivering an explicit counterexample to Cabello et al.'s claim and then identifying the shortcomings of their argument.

[44]Putting aside those views which would see it as law-like—nomological rather than ontological.(Dürr et al., 1995, 2012)

[45]In the multi-particle case, similarly, the motion of the occupied point in configuration space (representing the position of the n particles) is guided by the wavefunction assigned to the immediately surrounding region of configuration space.

[46]At least in approximations where coarse-graining means that a system—perhaps an isolated, or relatively isolated, degree of freedom—is being treated effectively as a finite-dimensional system.

3.3.6 Interlude: Landauer's Principle and Irreversibility

A great deal in this argument hinges on the application of Landauer's principle, often taken to be the claim that the implementation of a logically irreversible operation is accompanied by a dissipation[47] of $kT \ln 2$ units of heat per bit into the environment. An operation is considered to be logically irreversible, if the output of the operation does not uniquely determine the input (Landauer, 1961). Often, the concept of information is invoked in order to describe this logical irreversibility. Wiesner et al. (2012) write that "logically irreversible operations forget information about the computational device's preceding logical state." [p.4060].

To appreciate why such characterisations of Landauer's principle in terms of 'forgetting' can be misleading, however, we consider a logically irreversibly operation which can be implemented without any heat cost, the so-called RAND operation (Maroney, 2005). RAND randomises the logical state of a bit, regardless of its input state.

Physically, we may think of an implementation of RAND in the standard way of considering a molecule in a box, in which a partition is included. The molecule is originally on the left side (or right side) of the box and the whole system is in contact with a heat bath at temperature T_{HB}. Implementing RAND then simply requires one to remove the partition, wait for a sufficiently long time and then re-insert the partition. The operation is logically irreversible because the output (the randomly distributed molecule) does not uniquely determine the input (the molecule being in one of the two mutually exclusive states)—but obviously there was no heat exchange with the environment during this process. In a very naive sense, the RAND operation has 'erased information', but this erasure of information has taken place at no heat cost. Landauer's principle therefore cannot simply be the statement that

[47] Heat dissipation is generally taken to be thermodynamically irreversible.

the implementation of a logically irreversible operation leads to the dissipation of heat into the environment.

The most prominent application of Landauer's principle is therefore a rather special process—the so-called *Landauer erasure* process, a resetting operation that maps the state of a randomised bit back to some pre-defined initial state. For the above described molecule-in-a-box scenario, such an erasure can be implemented by removing the partition and then pushing it isothermally in from one side of the box, until the particle is found once again certainly in the left (or right) side of the box. For this last step, work is performed on the system and heat is transferred into the surrounding heat bath[48]. This is done in a thermodynamically reversible fashion and therefore corresponds to a *heat transfer* and not to a *heat dissipation*. Because the system ends up in a pre-defined state, the Landauer erasure is distinct from the RAND operation described above, although both of them are logically irreversible. In general, whether or not a given logical operation can be performed in a thermodynamically reversible or irreversible fashion depends on the choice of implementation. In principle any logical operation can be implemented in a thermodynamically reversible fashion. More details on this can be found in (Maroney, 2009b), whose approach we follow closely in this section[49].

What Landauer's principle in fact does is provide us with a link between logical operations and the fundamental microdynamics. Properly put, it states that a logical transformation, reversible or irreversible, must be accompanied by a minimal average heat dissipation into the environment according to

[48]Similarly in the quantum case for a qubit, one needs to step-wise raise one of the two energy levels to infinity (Barnett and Crutchfield, 2015).

[49]Another prominent account is that of Ladyman et al. (2007), who carefully analyse and defend Landauer's principle, but have a slightly different take on what is the domain of Landauer's principle, i.e. which logical operations are permissible.

$$\langle \Delta Q \rangle \geq -T_{HB} \Delta S, \tag{3.66}$$

where S_α and S_β refer to the states of the system and ΔS refers to the difference in von Neumann entropy given by

$$\Delta S = \sum_\beta P(\beta) \left(S_\beta - k \ln P(\beta) \right) - \sum_\alpha P(\alpha) \left(S_\alpha - k \ln P(\alpha) \right), \tag{3.67}$$

where α and β are the logical states involved in the transformation and ΔQ is the heat generated in the environment (Maroney, 2009b). Phrased like this, Landauer's principle not only becomes utterly unmysterious, it also becomes clear that whether or not heat is transferred into the environment solely depends on the von Neumann entropy of the system before and after the operation. This means in particular that one does not need to make any reference to information erasure or the like.

If we re-write Landauer's principle in a way that makes use of the concept of information by having it explicitly include the entropy of the information bearing degrees of freedom, the change in (Gibbs-) von Neumann entropy is related to the change in Shannon entropy by $\Delta S = \sum_\beta P(\beta) S_\beta - \sum_\alpha P(\alpha) S_\alpha + k \Delta H \ln 2$, with ΔH being the change in Shannon entropy. Making use of Equation (3.66), Landauer's principle then becomes

$$\Delta S_{NI} \geq -k \Delta H \ln 2, \tag{3.68}$$

where the entropy change of the non-information bearing degrees of freedom is given by the entropy change in the environment and the weighted entropy changes of the sub-ensembles: $\Delta S_{NI} = \Delta S_E + \sum_\beta P(\beta) S_\beta - \sum_\alpha P(\alpha) S_\alpha$. This quantity is increasing for logically irreversible, deterministic operations. For non-deterministic operations

however, it is decreasing, making Equation (3.66) more useful in the general context.

Now: Let us consider applying Landauer's principle in the form of Equation (3.66) to the repeatedly measured quantum system. It is clear that once the measurement process is up and running, the quantum system will be in a maximally mixed state at each time step, independent of the chosen measurement basis. This means that the difference in the von Neumann entropy of the quantum system before and after each measurement is zero. Given, furthermore, that the density matrix is the appropriate entity for calculating the entropy of a quantum system—even if one's interpretation involves further variables as is the case in the de Broglie–Bohm theory for example—it follows that Landauer's principle does not predict a heat cost for Type I interpretations: the lower bound on heat dissipation into the environment is zero[50] Notice that we have had to say nothing here of the storing or deleting or erasure of information.

But now, interestingly, we seem to have arrived at a contradiction: On the one hand, applying Landauer's principle to the successive measurements on a quantum system seems to entail a non-zero lower bound to the average heat cost per measurement. Whilst on the other it entails a zero lower bound; one which can be saturated. What needs to give?

In fact, nothing. This appearance of contradiction is misleading. The two results are not in fact in conflict. Both involve licit applications of Landauer's principle, but it is only the second (the zero heat minimum heat cost claim) which pertains to features of the quantum system.

[50]N.B. In the standard case in de Broglie-Bohm in which the distribution of particle positions is given by the Born rule then one will calculate the entropy of the system via its density matrix. If the distribution is not given by the Born-rule—the system is not in quantum equilibrium, in the phrase—there will be rather more pressing departures from standard quantum predictions than simply thermal ones. (Valentini, 1991, 2002)

To explain this point and to help clarify where Cabello et al.'s argument has gone wrong, we will now construct an explicit counterexample to their argument. It will become apparent that the heat cost they calculate does not stem from the quantum system itself but from the particular setup that is chosen. The heat cost will be shown to be due to *external* matters.

3.3.7 A Counterexample: Type I without Heat Dissipation

Spekkens' toy model (Spekkens, 2007) is explicitly taken to be a Type I interpretation. We will make use of this simple and transparent framework in order to illustrate how successive measurements in non-orthogonal bases do *not* lead to a predicted heat cost.

In Spekkens' toy model, measurements of the system only yield incomplete knowledge about the underlying state in such a way that the maximal amount of knowledge about the ontic state equals the lack of knowledge about it. A Spekkens qubit can be in one of four possible ontic states, '1', '2', '3' or '4'. Measurements are identified with questions of the form: is the system in state $1 \vee 2$ or $3 \vee 4$ (measurement in the $|0\rangle / |1\rangle$ basis), or, alternatively, $1 \vee 3$ or $2 \vee 4$ (measurement in the $|+\rangle / |-\rangle$ basis)? A measurement result of $|0\rangle$ will therefore yield a state of knowledge of $1 \vee 2$. Impressively, quantum behaviour of a single qubit for measurements restricted to the x, y, and z bases is fully recovered in this model.

If a system is prepared to be in $1 \vee 3$ but then measured to be in $1 \vee 2$, thereby performing the Spekkens equivalent of a non-orthogonal measurement, we know for a fact that the system must have been in ontic state 1 *before* the measurement (Spekkens, 2007). This means that it is possible to know the precise ontic state of the system at an earlier time, but impossible to know it at the current time. A

measurement in a non-orthogonal basis therefore *disturbs* the system in such a way that its values become non-definite in its previous basis.

Since we only consider measurements in the x and z directions, we can illustrate Spekkens' toy model conveniently by considering a classical particle entrapped in a two dimensional box in contact with a heat bath, as illustrated in Figure (3.3a). Each section of the box corresponds to one of the four ontic states the particle can be in. To perform measurements, either a horizontal or a vertical partition can be inserted. A measurement in the $|0\rangle / |1\rangle$ basis for example corresponds to inserting a partition *vertically* (Figure 3.3b)) and measuring on which side of the partition the particle is found. Performing a measurement in the $|+\rangle / |-\rangle$ basis in contrast corresponds to the insertion of the partition *horizontally* before measuring the the particle's position (Figure 3.3c)). There can only be one partition in the box at a given time, and so in order to perform a measurement in the $|+\rangle / |-\rangle$ basis by inserting the partition horizontally, we need to remove the partition from the previous measurement in the $|0\rangle / |1\rangle$ basis. This must happen sufficiently quickly compared to the free motion of the particle so as to ensure that two consecutive measurements in the same basis yield the same result. On the other hand, time between consecutive measurements must be sufficiently large for the particle's position to be uniformly distributed over the region to which it is confined. Alternatively, one could imagine that the measurement disturbs the particle in such a way, that its position is randomized over the restricted region.

With the above setup at hand, we now implement Cabello et al.'s experiment and include successive random measurements in non-orthogonal bases.

We begin by noting the need to distinguish between two notions of measurement which one might have in mind. On the one hand, one can consider some physical

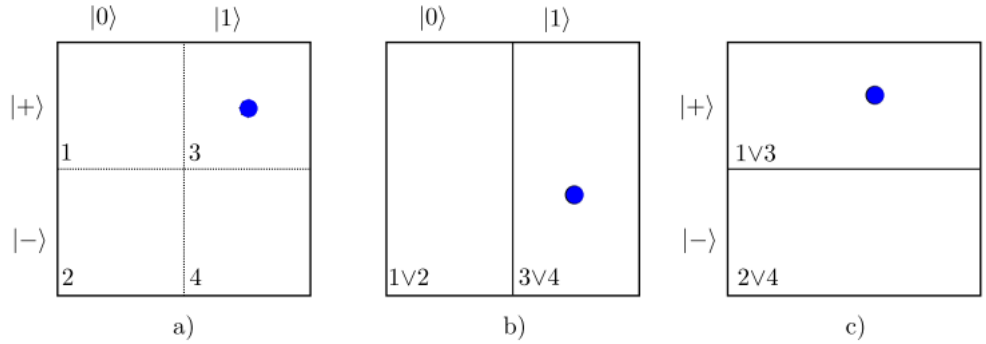

Figure 3.3: Illustration of standard measurements in Spekkens' toy model. We consider a classical particle in a two dimensional box. a) illustrates how at each instant, the particle is in one of 4 mutually exclusive ontic states. Measurements in the standard bases are modeled as the b) vertical or c) horizontal insertions of a partition into the box, followed by a location measurement that allows one to identify the particle's position as either being left/right or top/bottom. Two consecutive left/right and top/bottom measurements are non-commutative.

process taking place which leaves the system in some definite value of some observable. (A definite possessed value of the observable, whether or not anyone knows it.) On the other hand, one might consider an external observer or agent who comes to know what the definite value (or values) a system possesses at a given time are. (This difference is akin to the selective/non-selective measurement distinction familiar in quantum foundations.) Such an external observer need not be a person, but could be anything that reliably correlates with the measurement outcome, such as a memory cell.

We first consider the case in which there is no external agent. Measurements on the Spekkens particle are performed by either the horizontal or vertical insertion of the partition. In order to implement Cabello et al.'s experimental setup, we require that the orientation of the partition change randomly, in such a way that it has a probability of 50% of remaining in its previous position and a probability of 50% of changing its orientation from horizontal to vertical or vice versa. This

is an implementation of Cabello et al.'s random variable X_t. At each time step, the system will be in a well defined state with definite values. It is evident *prima facie* that there is no heat exchange with the environment at any point[51]. Moreover, this Spekkens setup is merely a more elaborate version of the previously introduced RAND operation. Evidently the system is in a definite state with definite values after each measurement, however, this "generation" of new values is not accompanied by a heat exchange with the environment.

One may object to the above reasoning by demanding that the partition *itself* must have a reset state, that it therefore must delete information and reset itself after each measurement. However, the described Spekkens' setup does not require a costly *Landauer erasure*: the system *overwrites* its previous state after each time stepAs there is no external agent who records the measurement outcomes by correlating a measurement apparatus with the system's state, there are no resources needed for this implementation of measurements.

The situation changes if we require the measurement performed and the result obtained to be determined and/or recorded by an external agent. Such an agent need not be human, but could be any system that is able to correlate itself with the Spekkens system and perform operations depending on the outcome. In the case of the experimental setup described above, a (memoryless) external agent needs to acquire at least two bits of information in order to determine which state the system is in at a given time t: one bit that determines whether the particle was measured in the x or z basis and one bit that determines the outcome of the measurement,

[51]Whether or not the act of removing and re-inserting the partition is thermodynamically reversible or irreversible is debatable given to the single particle nature of the experiment. One may argue that the removal of the partition resembles a free expansion, which is thermodynamically irreversible. However, if we take thermodynamic reversibility to be equal with the claim that the (Gibbs–) von Neumann entropy of the system remains constant, then the removal and re-insertion of the partition is indeed reversible, in accordance with Maroney (2009b). Either way, there is clearly no heat exchanged with the environment.

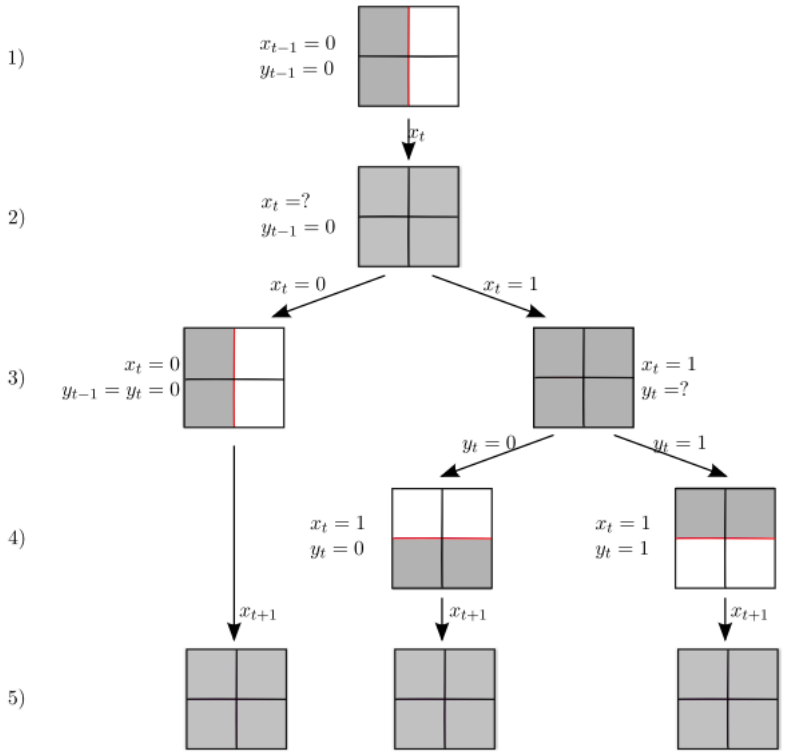

Figure 3.4: Illustration of the various steps in an exemplary measurement cycle. The system starts out in state $x_{t-1} = 0$ and $y_{t-1} = 0$, where x is the measurement basis (in this case left/right) and y the position of the particle (in this case left) (1). Then the random measurement in an unknown basis is performed (2). The system is now in a maximally mixed state as the particle's location is now fully unknown to an external observer. The previous measurement outcome y_{t-1} is kept for the case that the measurement basis x_t turns out to be the same as x_{t-1}. In the next step, the measurement basis x_t is determined by the external observer (3). If $x_t = x_{t-1}$, the position of the particle is known, as the previous measurement result has been kept and so only the time label changes, $y_{t-1} \to y_t$. If $x_t \neq x_{t-1}$, the external agent now knows the measurement basis, but still does not know the particle's location (whether top or bottom). To determine its location, the agent will need to perform a further measurement (4). In (5) the next random measurement is performed and the process is repeated.

namely whether the particle is left/right or top/bottom respectively. We imagine now a situation in which the agent is memoryless, i.e. has no access to the choice of measurement basis and the measurement result at time $t - 1$. If we allow the agent to have only two binary working-memory cells, one that reads out the choice

of basis given by X_t, and one that reads out the position of the particle, Y_t, then once the agent has determined the state of the system at time t, she needs to reset both memory cells in order to prepare herself for the next measurement cycle at time $t+1$. Given that resetting memory cells is costly, there will be a heat cost of $kT \ln 2$ associated with each measurement cycle.

This is where Cabello et al.'s result enters the picture. The described heat cost can in fact be reduced to 3/2 bits once we allow the agent to use a recording device, or a *memory*, which allows her to access the measurement results at the previous time-step[52] $t-1$. In this case, the agent does not need to perform the position measurement iff $x_t = x_{t-1}$, i.e. if she finds that the measurement basis at time t is the same as at time $t-1$. This is due to the fact that consecutive measurements in the same basis always yield the same measurement results. Given that she has access to the previous measurement result, she can therefore skip the location measurement in 50% of the cases and hence save resources. If the agent has access to a memory that specifies the previous measurement basis and measurement outcome, the average amount of bits needed to specify the measurement and outcome at time t can thereby be reduced to 3/2 bits. The minimal average heat cost per measurement cycle therefore becomes $\frac{3}{2}kT \ln 2$, the value Cabello et al. derived.

Figure 3.4 illustrates the various steps in an exemplary measurement cycle: after an input x_t, the state of the system has changed into a maximally mixed state. The first step of the agent must be to determine the measurement basis, i.e. the position of the partition. This step must be performed during *each* cycle, and so, given a finite memory for the agent and the need to create new blank memory states, there

[52] In practice, we need to grant the agent at least two more memory cells as resources, such as to be used as a memory for the measurement basis and measurement result at time $t-1$. In computational mechanics it is more common simply to supply the agent with an empty tape on which she records the measurement outcomes and at the same time allow her to access the tape on which the basis choices are written, thereby providing her with $\overleftarrow{y_t}$ and $\overleftarrow{x_t}$.

is always a heat cost of $kT \ln 2$ associated with this step. If $x_t = x_{t-1}$, nothing further happens, y_{t-1} changes into y_t, but no resources are required for this step and so the cycle is finished. If $x_t \neq x_{t-1}$ however, a position measurement must be performed that determines y_t. This once more leads to $kT \ln 2$ units of heat, since the agent only has limited memory and therefore needs to reset her memory before each measurement. Since this measurement of y_t is only performed half of the time, the average heat cost associated with it is only $\frac{1}{2}kT \ln 2$ units of heat. Adding up the various contributions leads to an average heat cost of $\frac{3}{2}kT \ln 2$ per measurement cycle, in accordance with Cabello et al.'s results. From the point of view of an external observer, however, the Spekkens quantum system itself is in a maximally mixed state from step 2) onwards.

The alleged quantum heat cost therefore merely results from the need to reset the various memory cells of the agent that are needed to record the measurement outcome.[53] The heat cost associated with Cabello et al.'s setup is thereby no more mysterious than the heat cost involved in the consecutive measurement of the outcomes of a fair coin flip, given limited resources: at each time step the measurement apparatus must be reset so as to be able to perform the consecutive measurements. Differently put, in terms of particles and boxes: performing consecutive RAND operations on a system leads to a heat cost iff the system's state is recorded at each time step, in which case the measurement apparatus needs to be reset before each measurement. What Cabello et al. have shown therefore, is simply that if we allow the agent to have a *memory*, the average heat cost for repeatedly recording the outcomes can be reduced from 2 bits to 3/2 bits, and no further.

Furthermore, in order to explain the origin of this heat cost, we did not need to

[53] One might compare with standard discussions of Maxwell's Demon (Maroney, 2009a) where the heat cost of resetting an agent's memory turns out to be rather important.

mention any *intrinsic values* which supposedly determine the future behaviour of the quantum system. The resources needed for Cabello et al.'s experiment are resources that are to be provided by the observer who wants to record the measurement results, and not by the quantum system itself. We note that the observer and his or her resources could well be purely classical; and we note moreover that the very same heat cost would be incurred even for a Type II interpretation.

3.3.8 Some Remarks on Computational Mechanics

The above examples demonstrate that the heat cost attributed to certain types of quantum interpretation by Cabello et al. leads back to an *external* agent repeatedly performing measurements with limited information storage available. The external agent does not need to be human but could equally be a machine and the calculated heat cost indeed gives a lower bound to the energy required to run such process. Cabello et al.'s result therefore clearly plays an important role for determining the resources needed to perform certain quantum computational tasks. What their argument does not establish however, is a distinctive and perhaps problematic heat cost for Type I interpretations. We can trace back the misinterpretation of their mathematical results to a lack of discrimination between the agent who acts and the system that is acted upon. Instead, agent and system together were treated as a unified computational machine, leading to the erroneous conclusion that the individual system itself was driving the computation and was the locus of the heat cost. This suggests that we need to be more careful when we apply computational mechanics to the foundations of quantum mechanics. In particular the tempting but somewhat misleading idea that information is a *concrete* rather than *abstract* entity, which gives it the status of a physical substance that can be transferred and

destroyed, leads to intuitions that must be double checked by contrasting them with real, physical situations (cf. (Timpson, 2013) for more details on the concrete vs. abstract distinction).

3.3.9 Conclusion

We have analysed the claim that there is a physical difference between two classes of quantum interpretations in terms of an excess heat generation in successive measurements, a difference which could in principle be tested experimentally. This is not so: there is no such differential heat production. In so far as there is a heat cost associated with successive randomly selected measurements on a quantum system, it arises from accounting for the external resources needed to record the performance and results of the measurement operations on the system, and not from a putative erasure process that takes place within the quantum system itself. That there is no heat cost involved with consecutive measurements of this kind arising from the quantum system itself—regardless of how one interprets probabilities—can immediately be seen upon calculating the difference of the von Neumann entropy of the quantum system between the various time steps, which difference is zero. This fact is reconciled with Cabello et al.'s quite correct mathematical result precisely by noting that the latter only concerns—when properly understood—costs of the external record, of the external agent.

The question of which interpretations might truly represent our world thereby remains unanswered by thermodynamical considerations concerning successive measurement scenarios.

Appendix

The quantity I_{erased} has its origin in the difference of the Shannon entropies at times t and $t-1$ as can be seen in the following, short derivation.

$$\begin{aligned} H\left(X_t Y_t S_t\right) - H\left(X_t Y_{t-1} S_{t-1}\right) &= H\left(X_t Y_t S_{t-1}\right) - H\left(X_t Y_{t-1} S_{t-1}\right) - H\left(S_{t-1}|X_t Y_t S_t\right) \\ &= H\left(X_t\right) + H\left(Y_t|S_t\right) + H\left(S_t\right) \\ &\quad - H\left(X_t\right) - H\left(Y_{t-1}|S_{t-1}\right) - H(S_{t-1}) - H\left(S_{t-1}|X_t Y_t S_t\right) \\ &= -H\left(S_{t-1}|X_t Y_t S_t\right) = -I_{erased}, \end{aligned}$$

where we assumed that the process is *unifilar*, namely $H\left(S_t|X_t Y_t S_{t-1}\right) = 0$, that the input is time-independent, i.e. $H\left(X_{t-1}\right) = H\left(X_t\right)$ and that the current causal state determines the output uniquely $H\left(Y_t|S_t\right) = 0$. Following Landauer's Principle, the physical implementation of a logical operation leads to a generation of heat equal to at least $-kT\ln 2$ times the total change in Shannon entropy. To calculate the total change, all involved random variables at a given time need to be taken into account. I_{erased} therefore provides a lower bound for the heat cost involved in the implementation of Cabello et al.'s experiment.

Conclusion

The aim of this thesis was to investigate the scope of thermodynamics, or in other words, the influence of its laws on the few and the many, the small and the large. I tried to give justice to Atkins' bold claim that thermodynamics is about 'almost everything', by surveying different parts of physics and examining the extent to which thermodynamics may be applied to those distinct areas. It was shown that the theory has an immense coverage, that indeed Einstein may have been right about the 'universal content' of thermodynamics. At the beginning of this thesis, I promised answers to the questions posed by Callender (2016). I want to honour this promise by now answering them in turn while at the same time summarising the results of this thesis.

"*What sub-systems of the universe does the second law govern?*"

Here it was shown that a large part of the answer depends on which second law we chose. The orthodox second law is very limited in its application as it is violated constantly on small scales, and so the answer here is "idealised macroscopic systems". For the most part, I was concerned with an amended version of the second law that prohibits the consistent and reliable conversion of heat into work. This latter understanding of thermodynamics opened the door for a much wider application. I showed in Chapter 1, that even thermodynamic systems that contain as little

as one particle can be thought of as *bona fide* thermodynamic objects, if they are describable by a set of robust thermodynamic parameters. In particular, I pointed out that the problems associated with one-molecule systems are generic problems about fluctuations and/or feedback processes which are merely simply amplified in the case of a system containing only a single molecule. Chapter 1 furthermore contained a discussion about the means-relative approach to thermodynamics, as advocated by Maxwell (1871), Gibbs (1878) and Jaynes (1965). I investigated various ways of how anthropocentrism could enter such a means-relative approach and came to the conclusion, that a means-relative understanding of heat, work and entropy does not imply an anthropocentric understanding thereof.

"*Are the principles of thermodynamics responsible for generalisations about black holes?*"

In Chapter 2 I investigated whether black holes are genuine thermodynamic objects with a thermodynamic entropy proportional to their horizon area. In addition to a discussion of the pertinent arguments for black hole entropy by Bekenstein (1973) and Hawking (1976), I furthermore challenged a recent criticism by Wüthrich (2017), in which he accused Bekenstein of establishing $S_{BH} = S_{TD}$ on the basis of information-theoretic considerations. I showed that Bekenstein himself never makes this claim, even though he does understand entropy in information-theoretic terms. The final part of the chapter was dedicated to establishing $S_{BH} = S_{TD}$ and was done using a box containing a black hole and a photon gas as the working medium for a Carnot cycle. It was shown that black holes indeed can be considered to have a thermodynamic entropy that scales with their horizon area.

"*What about the micro-realm?*"

Chapter 3 was dedicated to the micro-realm, or more precisely, to quantum mechanics. The first part of the chapter consisted of the analysis of a criticism expressed by Hemmo and Shenker (2006). The two authors claim that von Neumann's 1996 original argument, that introduces the von Neumann entropy as the quantum generalisation of the thermodynamic entropy, is flawed. This flaw allegedly disappears in the infinite particle limit but holds in the single and finite particle case. It was shown that the authors criticism a) is inconsistent with thermodynamics as it allows for the construction of a *perpetuum mobile* of the second kind, and b) boils down to the conventional Maxwell's demon debate in the context of Szilard engines. In particular it was shown that the behaviour of the joint von Neumann entropy of system and measurement apparatus mirrors the behaviour of the thermodynamic entropy.

The second part of this chapter derived the thermodynamic equations from quantum mechanics by making assumptions about the existence and properties of heat reservoirs. It was shown that no information-theoretic arguments need to be made in order to derive the von Neumann entropy as the quantum generalisation of the thermodynamic entropy. The derived relations were given in terms of expectation values, and I discussed how our interpretation of quantum mechanics influences our interpretation of heat, work and entropy. The subsequently reviewed so-called quantum resource theory of thermodynamics turned out to give a very general framework for thermodynamic-like behaviour, but required certain aspects to be put in by hand, such as the identification of the canonical states with the thermal states. The single-shot version of quantum mechanics was furthermore discussed briefly, and it was shown that for the creators of single-shot thermodynamics, the laws given in terms of expectation values are said to only hold in the thermodynamic limit. This resembles in a way the view of Hemmo and Shenker, encountered in the previous

section, who equally took the von Neumann entropy to only be identifiable with the thermodynamic entropy in the infinite particle limit.

The third and last part discussed a recent publication by Cabello et al., in which the authors claimed that thermodynamics allows for the derivation of an empirical difference between two classes of quantum interpretations. I provided a counterexample to their claim by making use of Spekkens classical toy model, and showed that the resulting heat cost in fact is fully accounted for by the recording device.

Given the richness of thermodynamics, many more questions remain that have been left unanswered. In this thesis, I was only concerned with the traditional derivations of thermodynamics by Carnot, Kelvin (1882), Clausius (1864) and Planck (1897). Axiomatic approaches such as the ones by Carathéodory (1909) and Lieb and Yngvason (1998), however, are viable alternatives. In particular their notion of 'adiabatic accessibility' that describes whether or not a thermodynamic state can be reached by another, given a set of operations on that sate, resembles very closely the resource theoretic approach. It will be interesting to examine whether these approaches can be applied to black holes, for example. This, however, will be part of another research project.

Bibliography

D. Z. Albert. *Time and Chance*. Harvard University Press, Cambridge, Mass., new ed edition edition, Feb. 2003.

P. Atkins. *Four Laws That Drive the Universe*. Oxford University Press, Oxford, UK, 2007.

P. Attard. *Thermodynamics and Statistical Mechanics: Equilibrium by Entropy Maximisation*. Elsevier Limited, San Diego, Calif, 2002.

L. E. Ballentine. The Statistical Interpretation of Quantum Mechanics. *Reviews of Modern Physics*, 42(4):358–381, Oct. 1970.

J. M. Bardeen, B. Carter, and S. W. Hawking. The Four Laws of Black Hole Mechanics. *Communications in Mathematical Physics*, 31(2):161–170, 1973.

N. Barnett and J. P. Crutchfield. Computational Mechanics of Input-Output Processes: Structured Transformations and the ϵ-Transducer. *Journal of Statistical Physics*, 161(2):404–451, Aug. 2015.

S. D. Bartlett, T. Rudolph, and R. W. Spekkens. Reference Frames, Superselection Rules, and Quantum Information. *Rev. Mod. Phys.*, 79:555–609, Apr 2007.

J. D. Bekenstein. Black Holes and Entropy. *Physical Review D*, 7(8):2333–2346, Apr. 1973.

J. D. Bekenstein. Black-Hole Thermodynamics. *Physics Today*, 33:24–31, Jan. 1980.

J. D. Bekenstein. Do We Understand Black Hole Entropy ? Sept. 1994. arXiv: gr-qc/9409015.

J. D. Bekenstein. Bekenstein-Hawking Entropy. *Scholarpedia*, 3(10):7375, Oct. 2008.

C. Bennett. Logical Reversibility of Computation. *IBM Journal of Research and Development*, 17(6):525–532, Nov. 1973.

D. Bohm. A Suggested Interpretation of the Quantum Theory in Terms of "Hidden" Variables. I. *Physical Review*, 85(2):166–179, Jan. 1952.

L. Boltzmann. Weitere Studien über das Wärmegleichgewicht unter Gasmolekülen. In *Kinetische Theorie II*, pages 115–225. Vieweg+Teubner Verlag, Wiesbaden, 1970.

L. Bombelli, R. K. Koul, J. Lee, and R. D. Sorkin. Quantum Source of Entropy for Black Holes. *Physical Review. D, Particles and Fields*, 34(2):373–383, July 1986.

F. Brandão, M. Horodecki, N. Ng, J. Oppenheim, and S. Wehner. The Second Laws of Quantum Thermodynamics. *Proceedings of the National Academy of Sciences*, 112(11):3275–3279, 2015.

F. G. Brandão and G. Gour. Reversible Framework for Quantum Resource Theories. *Physical Review Letters*, 115(7):070503, Aug. 2015.

F. G. S. L. Brandão, M. Horodecki, J. Oppenheim, J. M. Renes, and R. W. Spekkens. The Resource Theory of Quantum States Out of Thermal Equilibrium. *Physical Review Letters*, 111(25), Dec. 2013.

J. Bricmont. Science of Chaos or Chaos of Science? *Annals of the New York Academy of Sciences*, 775(1):131–175, June 1995.

P. W. Bridgman. *The Nature of Thermodynamics*. Harvard University Press, Cambridge, Mass., Jan. 1941.

H. Brown, W. Myrvold, and J. Uffink. Boltzmann's H- Theorem, its Discontents, and the Birth of Statistical Mechanics. *Studies in History and Philosophy of Science Part B - Studies in History and Philosophy of Modern Physics*, 40(2):174–191, 5 2009.

H. R. Brown. The Reality of the Wavefunction: Old Arguments and New, Apr. 2017. URL `http://philsci-archive.pitt.edu/12978/`.

H. R. Brown and J. Uffink. The Origins of Time-Asymmetry in Thermodynamics: The Minus First Law. In *Studies In History and Philosophy of Modern Physics*, pages 525–538, 2001.

A. Cabello, M. Gu, O. Gühne, J.-A. Larsson, and K. Wiesner. Thermodynamical Cost of Some Interpretations of Quantum Theory. *Physical Review A*, 94(5):052127, Nov. 2016.

C. Callender. Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy. *Journal of Philosophy*, 96(7):348–373, 1999.

C. Callender. Thermodynamic Asymmetry in Time. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, winter 2016 edition, 2016.

C. Carathéodory. Untersuchungen über die Grundlagen der Thermodynamik. *Mathematische Annalen*, 67:355–386, 1909.

S. Carnot. Réflexions sur la puissance motrice du feu et sur les machines propres à développer cette puissance. In *Annales scientifiques de l'École Normale Supérieure*, volume 1, pages 393–457, 1872.

C. M. Caves, C. A. Fuchs, and R. Schack. Quantum Probabilities as Bayesian Probabilities. *Physical Review A*, 65(2):022305, Jan. 2002.

C. M. Caves, C. A. Fuchs, and R. Schack. Subjective Probability and Quantum Certainty. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 38(2):255–274, 2007.

P. J. Clark. Statistical Mechanics and the Propensity Interpretation of Probability. Lecture Notes in Physics, pages 271–281. Springer, Berlin, Heidelberg, 2001.

R. Clausius. *Abhandlungen über die mechanische Wärmetheorie*, volume 1. F. Vieweg und Sohn, Braunschweig, Ger, 1864.

R. Clausius. *Die mechanische Wärmetheorie*, volume 1. F. Vieweg und Sohn, Braunschweig, 3 edition, 1887.

J. P. Crutchfield and K. Young. Inferring statistical complexity. *Physical Review Letters*, 63(2):105–108, July 1989.

E. Curiel. Classical Black Holes Are Hot. *arXiv preprint arXiv:1408.3691*, 2014.

P. S. Custodio and J. E. Horvath. Thermodynamics of Black Holes in a Finite Box. *American Journal of Physics*, 71(12):1237–1241, Dec. 2003.

O. Dahlsten, R. Renner, E. Rieper, and V. Vedral. Inadequacy of von Neumann Entropy for Characterizing Extractable Work. *New Journal of Physics*, 13(05), 2011.

L. del Rio, L. Kraemer, and R. Renner. Resource Theories of Knowledge. *arXiv preprint arXiv:1511.08818*, 2015.

K. G. Denbigh and J. S. Denbigh. *Entropy in Relation to Incomplete Knowledge*. Cambridge University Press, Cambridge, UK, 1985.

D. Deutsch. Quantum Theory of Probability and Decisions. *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, 455(1988): 3129–3137, Aug. 1999.

F. Dizadji-Bahmani, R. Frigg, and S. Hartmann. Who's Afraid of Nagelian Reduction? *Erkenntnis (1975-)*, 73(3):393–412, 2010.

J. Dougherty and C. Callender. Black Hole Thermodynamics: More Than an Analogy? Oct. 2016. URL `http://philsci-archive.pitt.edu/13195/`.

D. Dürr, S. Goldstein, and N. Zanghi. Bohmian Mechanics and the Meaning of the Wave Function. *arXiv preprint quant-ph/9512031*, Dec. 1995.

D. Dürr, S. Goldstein, and N. Zanghi. *Quantum Physics Without Quantum Philosophy*. Springer Science and Business Media, Nov. 2012.

S. A. Eddington. *The Nature of the Physical World*. J.M. Dent & Sons, London, 1935.

P. Ehrenfest and T. Ehrenfest. *Begriffliche Grundlagen der statistischen Auffassung in der Mechanik*. Druck und Verlag von B. G. Teubner, 1909.

A. Einstein. Beiträge zur Quantentheorie. *Verh. d. deutsch. phys. Gesellschaft*, 12, 1914.

A. Einstein. as quoted in M.J. Klein,"Thermodynamics in Einstein's Universe". *Science*, 157, 1967.

N. Erez, G. Gordon, M. Nest, and G. Kurizki. Thermodynamic Control by Frequent Quantum Measurements. *Nature*, 452(7188):724–727, 2008.

H. Everett. "Relative State" Formulation of Quantum Mechanics. *Reviews of Modern Physics*, 29(3):454–462, July 1957.

E. Fermi. *Thermodynamics*. Courier Corporation, June 1956.

R. H. Fowler and E. Guggenheim. *Statistical Thermodynamics:*. Cambridge University Press, Cambridge, UK, 1 edition, 1956.

R. Frigg. Typicality and the Approach to Equilibrium in Boltzmannian Statistical Mechanics. *Philosophy of Science*, 76(5):997–1008, 2009.

R. Frigg and C. Hoefer. Probability in GRW Theory. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 38(2):371–389, 2007.

R. Frigg and C. Werndl. Entropy-A Guide for the Perplexed. In C. Beisbart and S. Hartmann, editors, *Probabilities in Physics*. Oxford University Press, Cambridge, UK, 2011.

V. Frolov and I. Novikov. *Black Hole Physics: Basic Concepts and New Developments*. Springer, Dordrecht ; Boston, 1998 edition, Nov. 1998.

C. Fuchs, N. David Mermin, and R. Schack. An Introduction to QBism with an Application to the Locality of Quantum Mechanics. *American Journal of Physics*, 82, Nov. 2013.

C. A. Fuchs. Quantum Mechanics as Quantum Information (and only a little more). *arXiv preprint arXiv:quant-ph/0205039*, May 2002.

C. A. Fuchs and A. Peres. Quantum Theory Needs No 'Interpretation'. *Physics Today*, 53(3):70–71, Jan. 2007.

A. J. P. Garner, J. Thompson, V. Vedral, and M. Gu. When is Simpler Thermodynamically Better? *arXiv preprint arXiv:1510.00010*, Sept. 2015.

J. Gemmer, M. Michel, and G. Mahler. *Quantum Thermodynamics: Emergence of Thermodynamic Behavior Within Composite Quantum Systems*. Springer Science & Business Media, 2009.

G. C. Ghirardi, A. Rimini, and T. Weber. Unified Dynamics for Microscopic and Macroscopic Systems. *Physical Review D*, 34(2):470–491, July 1986.

D. C. Giancoli. *Physik: Lehr- und Uebungsbuch*. Pearson Deutschland GmbH, 2010.

J. W. Gibbs. On the Equilibrium of Heterogeneous Substances. *American Journal of Science*, Series 3 Vol. 16(96):441–458, Dec. 1878.

S. Goldstein. Boltzmann's Approach to Statistical Mechanics. In J. Bricmont, G. Ghirardi, D. Duerr, F. Petruccione, M. C. Galavotti, and N. Zanghi, editors, *Chance in Physics*, number 574 in Lecture Notes in Physics, pages 39–54. Springer Berlin Heidelberg, 2001.

S. Goldstein. Bohmian Mechanics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, summer 2017 edition, 2017.

J. Goold, M. Huber, A. Riera, L. d. Rio, and P. Skrzypczyk. The Role of Quantum Information in Thermodynamics – A Topical Review. *Journal of Physics A: Mathematical and Theoretical*, 49(14):143001, 2016.

G. Gour, M. P. Müller, V. Narasimhachar, R. W. Spekkens, and N. Yunger Halpern. The resource theory of informational nonequilibrium in thermodynamics. *Physics Reports*, 583:1–58, July 2015.

H. Greaves. Probability in the Everett Interpretation. *Philosophy Compass*, 2(1): 109–128, Jan. 2007.

A. Greven, G. Keller, and G. Warnecke. *Entropy*. Princeton University Press, Sept. 2014.

N. Harrigan and R. W. Spekkens. Einstein, Incompleteness, and the Epistemic View of Quantum States. *Foundations of Physics*, 40(2):125–157, 2010.

S. W. Hawking. Particle Creation by Black Holes. *Communications in Mathematical Physics*, 43(3):199–220, 1975.

S. W. Hawking. Black Holes and Thermodynamics. *Physical Review D*, 13(2): 191–197, Jan. 1976.

S. W. Hawking and D. N. Page. Thermodynamics of Black Holes in Anti-de Sitter Space. *Communications in Mathematical Physics*, 87(4):577–588, Dec. 1983.

M. Hemmo and O. Shenker. Von Neumann's Entropy Does Not Correspond to Thermodynamic Entropy. *Philosophy of Science*, 73(2):153–174, 2006.

L. Henderson. The Von Neumann Entropy: A Reply to Shenker. *The British Journal for the Philosophy of Science*, 54(2):291–296, 2003.

M. Horodecki and J. Oppenheim. Fundamental Limitations for Quantum and Nanoscale thermodynamics. *Nature Communications*, 4:2059, June 2013. doi: 10.1038/ncomms3059.

R. Horodecki, P. Horodecki, M. Horodecki, and K. Horodecki. Quantum Entanglement. *Rev. Mod. Phys.*, 81:865–942, Jun 2009.

P. Humphreys. Why Propensities Cannot be Probabilities. *The Philosophical Review*, 94(4):557–570, 1985.

E. T. Jaynes. Information Theory and Statistical Mechanics. *Physical Review*, 106 (4):620–630, May 1957.

E. T. Jaynes. Gibbs vs Boltzmann Entropies. *American Journal of Physics*, 33: 391–398, May 1965.

Kelvin. *Mathematical and Physical Papers*, volume 1. Cambridge University Press, Cambridge, UK, 1882.

A. G. Kofman and G. Kurizki. Frequent Observations Accelerate Decay: The Anti-Zeno effect. *Zeitschrift für Naturforschung A*, 56(1-2), 2001.

J. Ladyman, S. Presnell, A. J. Short, and B. Groisman. The Connection Between Logical and Thermodynamic Irreversibility. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 38:58–79, Mar 2007.

J. Ladyman, S. Presnell, and A. J. Short. 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, 2008.

R. Landauer. Irreversibility and Heat Generation in the Computing Process. *IBM journal of research and development*, 5(3):183–191, 1961.

D. Lavis. Boltzmann and Gibbs: An Attempted Reconciliation. *Studies In History and Philosophy of Science Part B: Studies In History and Philosophy of Modern Physics*, 36:245–273, Feb. 2004.

D. Lavis. Boltzmann, Gibbs and the Concept of Equilibrium. *Philosophy of Science*, 75, Nov. 2007.

D. Lavis. An Objectivist Account of Probabilities in Statistical Physics. page 51. Oxford University Press, Oxford, UK, 2011.

J. L. Lebowitz. Macroscopic Laws, Microscopic Dynamics, Time's Arrow and Boltzmann's entropy. *Physica A: Statistical Mechanics and its Applications*, 194 (1):1–27, Mar. 1993.

A. Lenard. Thermodynamical Proof of the Gibbs Formula for Elementary Quantum Systems. *Journal of Statistical Physics*, 19(6):575–586, Dec 1978.

E. H. Lieb and J. Yngvason. A Guide to Entropy and the Second Law of Thermodynamics. In P. B. Nachtergaele, P. J. P. Solovej, and P. J. Yngvason, editors, *Statistical Mechanics*, pages 353–363. Springer Berlin Heidelberg, 1998.

S. Lloyd. Quantum Thermodynamics: Excuse our ignorance. *Nature Physics*, 2: 727–728, Nov. 2006.

O. Lombardi and D. Dieks. Modal Interpretations of Quantum Mechanics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Winter edition, 2016.

O. Maroney. Information Processing and Thermodynamic Entropy. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Fall edition, 2009a.

O. J. E. Maroney. The (Absence of a) Relationship Between Thermodynamic and Logical Reversibility. *Studies in History and Philosophy of Science Part B*, 36(2): 355–374, 2005.

O. J. E. Maroney. The Physical Basis of the Gibbs-von Neumann entropy. *arXiv preprint arXiv:quant-ph/0701127*, 2007.

O. J. E. Maroney. Generalising Landauer's Principle. *Physical Review E*, 79(3), Mar. 2009b.

J. Maxwell. Illustrations of the dynamical theory of gases. Part I. On the motions and collisions of perfectly elastic spheres. *The London, Edinburgh, and Dublin Philosophical Magazine and Journal of Science*, 19(4):19–32, 1860.

J. Maxwell. Diffusion. In *Encyclopedia Britannica*, volume 7, pages 214–221. Reprinted in Niven (1965, pp.625-646), 9 edition, 1878.

J. C. Maxwell. *Theory of Heat*. Longmans, 1871.

J. C. Maxwell. *The Scientific Papers*. Dover Publications, Mineola, NY, 1965.

J. C. Maxwell. *Maxwell on Heat and Statistical Mechanics: On "Avoiding All Personal Enquiries" of Molecules*. Lehigh University Press, 1995.

C. W. Misner, K. S. Thorne, and J. A. Wheeler. *Gravitation*. W. H. Freeman, San Francisco, first edition, Jan. 1973.

B. Misra and E. C. G. Sudarshan. The Zeno's Paradox in Quantum Theory. *Journal of Mathematical Physics*, 18(4):756–763, 1977.

W. C. Myrvold. 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, 2011.

P. Nag. *Basic & Applied Thermodynamics*. McGraw Hill Education, New Delhi; Singapore, second edition, July 2017.

J. D. Norton. Waiting for Landauer. *Studies in History and Philosophy of Science Part B*, 42(3):184–198, 2011.

T. Opatrný and L. Richterek. Black Hole Heat Engine. *American Journal of Physics*, 80(1):66–71, Dec. 2011.

D. N. Page. Particle Emission Rates from a Black Hole: Massless particles from an uncharged, nonrotating hole. *Physical Review D*, 13(2):198–206, Jan. 1976.

M. H. Partovi. Quantum Thermodynamics. *Physics Letters A*, 137:440–444, Jun 1989.

O. Penrose. *Foundations of Statistical Mechanics*. Pergamon, first edition, 1970.

M. Planck. *Vorlesungen über Thermodynamik*. Leipzig, Veit & comp., 1897.

M. Planck. *Treatise on Thermodynamics*. Dover Publications Inc., New York, 7th revised edition edition edition, Jan. 1991.

S. Popescu, A. J. Short, and A. Winter. The Foundations of Statistical Mechanics from Entanglement: Individual States vs. Averages. *arXiv preprint arXiv:quant-ph/0511225*, Nov. 2005.

K. Popper. *The Logic of Scientific Discovery*. Routledge, second edition, Feb. 2002.

K. R. Popper. The Propensity Interpretation of the Calculus of Probability, and the Quantum Theory. pages 65–70. Butterworths, 1957.

K. R. Popper. The Propensity Interpretation of Probability. *British Journal for the Philosophy of Science*, 10(37):25–42, 1959.

C. E. A. Prunkl and C. G. Timpson. On the Thermodynamical Cost of Some Interpretations of Quantum Theory. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, Jan. 2018.

W. Pusz and S. L. Woronowicz. Passive States and KMS States for General Quantum systems. *Communications in Mathematical Physics*, 58(3):273–290, 1978.

H. Reichenbach. *The Theory of Probability*. University of California Press, 1971.

L. d. Rio, J. Aberg, R. Renner, O. Dahlsten, and V. Vedral. The Thermodynamic Meaning of Negative Entropy. *Nature*, 474(7349):61–63, June 2011.

C. Rovelli. Relational Quantum Mechanics. *International Journal of Theoretical Physics*, 35(8):1637–1678, Aug. 1996.

T. Sagawa and M. Ueda. Second Law of Thermodynamics with Discrete Quantum Feedback Control. *Physical Review Letters*, 100(8):080403, Feb. 2008.

S. Saunders. Time, Quantum Mechanics, and Probability. *Synthese*, 114(3):373–404, 1998.

S. Saunders. The Gibbs Paradox. *Entropy*, 20(8):552, 2018.

L. S. Schulman and B. Gaveau. Ratcheting up Energy by Means of Measurement. *Physical Review Letters*, 97(24):240405, 2006.

G. L. Sewell. Stability, Equilibrium and Metastability in Statistical Mechanics. *Physics Reports*, 57(5):307–342, Jan 1980.

O. R. Shenker. Is -kTr(ρlnρ) the Entropy in Quantum Mechanics. *British Journal for the Philosophy of Science*, 50(1):33–48, 1999.

L. Sklar. *Physics and Chance: Philosophical Issues In The Foundations Of Statistical Mechanics*. Cambridge University Press, Cambridge, Dec. 1995.

L. Sklar. The Reduction(?) Of Thermodynamics to Statistical Mechanics. *Philosophical Studies*, 95(1-2):187–202, 1999.

P. Skrzypczyk, A. J. Short, and S. Popescu. Work Extraction and Thermodynamics for Individual Quantum Systems. *Nature Communications*, 5:4185, June 2014.

M. v. Smoluchowski. Gültigkeitsgrenzen des zweiten Hauptsatzes der Wärmetheorie. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 22:61–64, 1913.

A. Sommerfeld. *Vorlesungen Über Theoretische Physik, Bd.5, Thermodynamik und Statistik*. Deutsch, Thun, Jan. 1988.

R. W. Spekkens. Evidence for the epistemic view of quantum states: A toy theory. *Physical Review A*, 75(3):032110, Mar. 2007.

A. Strominger and C. Vafa. Microscopic Origin of the Bekenstein-Hawking Entropy. *Physics Letters B*, 379(1):99–104, June 1996.

L. Susskind. Some Speculations about Black Hole Entropy in String Theory. *arXiv preprint arXiv:hep-th/9309145*, Sept. 1993.

L. Szilard. Über die Ausdehnung der phänomenologischen Thermodynamik auf die Schwankungserscheinungen. *Zeitschrift für Physik*, 32(1):753–788, Dec 1925.

L. Szilard. On the Decrease of Entropy in a Thermodynamic System by the Intervention of Intelligent Beings', Szilard (1972). *Reprinted in Leff and Rex (1990)*, pages 124–133, 1929.

W. Thirring. Systems with Negative Specific Heat. *Zeitschrift für Physik A Hadrons and Nuclei*, 235(4):339–352, Aug. 1970.

C. G. Timpson. Quantum Bayesianism: A study. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 39(3): 579–609, Sept. 2008.

C. G. Timpson. *Quantum Information Theory and the Foundations of Quantum Mechanics*. Oxford University Press, Oxford, Apr. 2013.

R. C. Tolman. *The Principles Of Statistical Mechanics*. Clarendon Press, Oxford, 1938.

M. Tribus and E. C. McIrvine. Energy and Information. *Scientific American*, 225(3): 179–190, 1971.

J. Uffink. Bluff Your Way in the Second Law of Thermodynamics. *Studies in History and Philosophy of Science Part B*, 32(3):305–394, 2001.

L. Vaidman. Probability in the Many-Worlds Interpretation of Quantum Mechanics. pages 299–311. Springer, 2012.

L. Vaidman. Many-Worlds Interpretation of Quantum Mechanics. In E. N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, fall edition, 2016.

A. Valentini. Signal-locality, Uncertainty, and the Subquantum H-Theorem. II. *Physics Letters A*, 158(1):1–8, Aug. 1991.

A. Valentini. Subquantum Information and Computation. *Pramana*, 59(2):269–277, Aug. 2002.

R. von Mises. *Wahrscheinlichkeit, Statistik und Wahrheit: Einführung in d. neue Wahrscheinlichkeitslehre u. ihre Anwendung*. Schriften zur wissenschaftlichen Weltauffassung. Springer-Verlag, Berlin Heidelberg, second edition, 1936.

J. von Neumann. *Mathematische Grundlagen der Quantenmechanik*. Springer, second edition, 1996.

R. M. Wald. Black Holes and Thermodynamics. In Z. Z. Venzo De Sabbata, editor, *Black Hole Physics*, NATO ASI Series, pages 55–97. Springer, Dordrecht, 1992.

R. M. Wald. Black Hole Entropy is Noether charge. *Physical Review. D, Particles and Fields*, 48(8):R3427–R3431, Oct. 1993.

R. M. Wald. The Thermodynamics of Black Holes. *Living Rev. Relativity*, 4, 2001.

D. Wallace. Everettian rationality: Defending Deutsch's approach to probability in the Everett interpretation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 34(3):415–439, Sept. 2003.

D. Wallace. *Emergent Multiverse: Quantum Theory According to the Everett Interpretation*. Oxford University Press, Oxford, U.K, July 2012.

D. Wallace. Inferential vs. Dynamical Conceptions of Physics. *arXiv preprint arXiv:1306.4907*, 2013a.

D. Wallace. *The Non-Problem of Gibbs vs. Boltzmann Entropies*. manuscript, 2013b.

D. Wallace. Thermodynamics as Control Theory. *Entropy*, 16(2):699–725, Jan. 2014.

D. Wallace. The Case for Black Hole Thermodynamics, Part I: phenomenological thermodynamics. *arXiv preprint arXiv:1710.02724*, Oct. 2017.

C. Weedbrook, S. Pirandola, R. García-Patrón, N. J. Cerf, T. C. Ralph, J. H. Shapiro, and S. Lloyd. Gaussian Quantum Information. *Rev. Mod. Phys.*, 84:621–669, May 2012.

A. Wehrl. General Properties of Entropy. *Reviews of Modern Physics*, 50(2):221–260, Apr 1978.

C. Werndl and R. Frigg. 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, Feb. 2015.

J. A. Wheeler and W. H. Zurek. *Quantum Theory and Measurement*. Princeton University Press, Princeton, NJ, July 2014.

K. Wiesner, M. Gu, E. Rieper, and V. Vedral. Information-Theoretic Lower Bound on Energy Cost of Stochastic Computation. In *Proceedings of the Royal Society of London A: Mathematical, Physical and Engineering Sciences*, volume 468, pages 4058–4066. The Royal Society, 2012.

C. Wüthrich. Are Black Holes About Information? *arXiv preprint arXiv:1708.05631*, Aug. 2017.

E. Zermelo. Über einen Satz der Dynamik und die mechanische Wärmetheorie. *Annalen der Physik*, 293(3):485–494, 1896.