THE CONTINGENT LAW
A tale of Maxwell's Demon

Victor Gijsbers

October 9, 2004

Contents

Overview

Aims of the thesis

This thesis is on Maxwell's Demon, the tiny and neat-fingered being that can break the second law of thermodynamics at will. It was introduced by Maxwell in 1867, in an attempt to shed light on the second law of thermodynamics in the framework of statistical physics. Now, more than a century after its summoning, the demon receives as much scholarly attention as ever. The impressive amount of recent publications (see for instance the chronological bibliography in Leff and Rex 2003, [20]) shows that the demon unabatedly creates controversy wherever it goes. In addition, it creates confusion. Neither the aims of the debates surrounding it, nor the rules by which success and failure in these debates ought to be judged, are clear. This has been pointed out in considerable detail by Earman and Norton 1998 & 1999 ([10], [11]).

One major aim of the present thesis has been to point out clearly what the major questions are, how they should be approached and what role the demon ought to play. My ideas concerning this are unfolded in the first part, **The contingent second law**. Building on the basis thus laid, I argue that the second law is contingent on some very specific features of our world: basically (but not merely) on the fact there are no particles smaller than atoms which can be used to build solid structures. The non-existence of Maxwell's Demon therefore cannot be proved on the basis of general principles.

Yet this is exactly what many schools of thought in the debate surrounding the demon claim to have done, either on the basis of fluctuations, or of measurement, or information erasure. If the conclusions I reached in the first part are correct, these claims must be incorrect. The second part of the the thesis, **Tales of the exorcists**, tries to pinpoint the mistakes made in these approaches by uncovering the hidden assumptions their proponents have made. Most of this discussion revolves about the question whether there is a deep connection between entropy and information. Special attention is devoted to the Landauer-Bennett scheme, as it has been at the forefront of recent discussions.

Because the debates surrounding Maxwell's Demon are so complex and often opaque, I have deemed it unprofitable to address specific positions in the literature in the first part of the thesis. I felt that it was necessary to approach the subject in as clear and straight a line as possible, without making detours to complex issues full of poorly defined terms or what I deem misconceptions. Such an approach enables me to create an image of the Maxwell's Demon problem that is relatively unproblematic, and from which I can approach the problems surfacing in the exorcist literature. One needs a perspective before one can look; in the first part of the thesis I create a perspective, in the second part I use it as a vantage point to survey and judge existing debates.

The goals of the thesis are fourfold. *First*, to clarify the questions which Maxwell's Demon may help us ask and answer, and to detail the way in which such an answer might be reached. *Second*, to establish the claim that the second law does not hold on any 'general' principles, but as a result of very specific features of our physics. *Third*, to explain how the schools of exorcism came to the conclusion that the second law could be proved on general principles, even though this is false. *Fourth*, to argue that there is no deep connection between entropy and information. A more detailed chapter-by-chapter description of the

thesis's structure is given below.

Chapter-by-chapter overview

- In the **Prologue: Birth of a Demon**, a playful introduction to the second law of thermodynamics and the properties of Maxwell's Demon is presented. This prologue can be skipped by anyone with rudimentary knowledge of the demon.

- **Chapter 1: The question of contingency** discusses two strange features of the second law. The first is its statistical nature, the second the unclarity surrounding the question which laws and features of the world it is contingent on. Then the role of the demon as a teacher on these two subjects is scrutinised, and the conclusion reached that it can only clarify the latter issue.

- In **Chapter 2: Portrait of a Demon** a phenomenal definition of Maxwell's Demon is developed. Attention is given to the question why and in what manner the demon ought to operate in a cycle, as well as to the use of ensembles and probabilities needed to characterise a successful demon. Various definitions of entropy are presented as candidates for giving a quantitative definition of Maxwell's Demon, but are discarded in favour of the phenomenal definition.

- **Chapter 3: An argument for necessity** discusses not only what I believe to be the most appealing argument for the second law's necessity, but also a counter-argument which defeats it and shows – so I claim – that the second law cannot be proved from classical mechanics. In order to discuss these issues, the chapter starts with an investigation into the ways in which thermodynamical quantities such as heat and work are represented in models of classical mechanics.

- **Chapter 4: Matters of scale and contingency** argues that once the microscopic/macroscopic dichotomy has been abandoned, we can see the importance of matters of scale for the second law. It is shown that the second law is contingent on facts that explicitly deal with the existence of natural scales and the relative abundance of objects of certain sizes and properties. It is thus proven that the second law is contingent.

- **Interlude: An imaginary history** sketches the history of exorcism as perceived by recent exorcists: a succession of better and more insightful attempts to banish the demon, culminating in the Landauer-Bennett school.

- In **Chapter 5: Doomed by fluctuations** the first stage of exorcism is discussed, including the well-known trapdoor of Smoluchowski and the ratchet and pawl of Feynman. It is argued that the fluctuation-exorcisms are based upon the 'fundamental assumption' that all parts of a system can be assigned a comparable temperature. An extended version of this assumption is the claim that all parts of a system must be described by the canonical distribution function.

- In **Chapter 6: The details of measurement**, Szilard's suggestion that measurements increase entropy and Brillouin's arguments that this defeats Maxwell's Demon are inspected. It is found that there is a sound core in these ideas, which can serve as a valuable addition to the fluctuation-approach, from which it does not differ dramatically. Brillouin's attempt to defeat the demon by arguing that entropy and information are related in a fundamental way is seen to fail.

- **Chapter 7: Erasure: the new paradigm** treats of the suggestion by Rolf Landauer and Charles Bennett that information erasure always has an entropy cost, and that this defeats the demon. We find two classes of proofs: those that use the extended fundamental assumption, and are therefore not very good exorcisms; and those that try to prove a deeper link between exorcism and entropy. The latter are found to be surrounded by confusion, but the valid core is exactly the State Space Contraction argument of chapter 3. The conclusion is that the Landauer-Bennett paradigm is not an especially successful mode of exorcism.

- In **Epilogue: Return of the Demon**, another playful aside, the Demon returns to earth to haunt the dreams of physicists once more.

- Finally, the **Conclusion** summarises the main points of this thesis, and ends with a happy affirmation of the demon's continued well-being.

Prologue: Birth of a Demon

Scene I: The second law of thermodynamics

A sunny park in Heaven. James Clark Maxwell is sitting on a bench, his eyes closed, enjoying the fine weather of the afterlife. He is approached by a somewhat worried looking Saint Peter.

Saint Peter James, I hope I'm not coming at an inopportune moment. There is a problem we truly need your help with.

Maxwell Opens his eyes. Good morning, Peter! You can call on me any time of the celestial day, and your pleas will never go unanswered. Although I must confess that I cannot imagine how saints or angels can have troubles *I* can help with.

Saint Peter We need to call on your expertise and knowledge, wise friend. Do you remember the demon you once summoned to earth? No, never mind my stupid question; of course you do.

Maxwell Most certainly; my tiny friend and I have had many interesting discussions after it was forced to leave Earth in the early 20^{th} century. That hurt me, it did – the cute little creature rejected and exorcised by the physical community. It was like a son to me. But they did not understand it, they feared it and thought it was an enemy to be combatted.

Saint Peter The reason I'm calling on your knowledge is that the demon has applied with the Celestial Office for readmittance to Earth. Its possible existence has been the centre of heated debate in a part of the physical community for quite some time now, and according to the demon's application form the strength of the exorcisms is on the wane. Enough so, it says, that it should be permitted to return to Earth. We are trying to evaluate this claim.

Maxwell I hope you will not ask me to decide the issue – I'm not even remotely impartial! The demon simply belongs on Earth. I still think its very purpose is to teach humanity about the second law.

Saint Peter What I would like you to do, my friend, is tell me all about the demon. What is it like? How and why did you summon it? We lack knowledge at the Celestial Office, and I need you to educate me.

Maxwell You want me to tell the story of the demon? Good, but we'll have to start at the beginning, decades before my summoning.

Saint Peter You have my complete attention.

Maxwell Good. I'll first explain the second law of thermodynamics to you. The latter half of the nineteenth century saw both the rise of thermodynamics as an important subdiscipline of physics, and its gradual and partial replacement by the later theory of statistical mechanics. Thermodynamics was set up as a theory that could describe quantities of macroscopic systems in equilibrium such as temperature, heat and pressure and predict the behaviour of gasses and other substances when heated, cooled, compressed, expanded, stirred, and so forth. All of this was to be accomplished without any theorising on the constitution of matter: thus, the scientific results would not be based on such highly dubious and debatable assumptions as the atomic theory.

Saint Peter But matter *does* consist of atoms, does it not?

Maxwell That is what we believe nowadays, but in the 19^{th} century many people were suspicious of the atomic theory. In any case, the results of thermodynamical research were many, the most important of which were summarised in the two main 'laws of thermodynamics'. The first law asserted the conservation of energy; the second the non-decrease of the rather abstract quantity called 'entropy'. This second law of thermodynamics was meant to be the expression of the world's irreversibility, of the lamentable fact that energy always degrades from a form with which we can do work to a form with which we can't.

Saint Peter 'Degradation of energy' – that sounds like an ominous prelude to the End of Times! What is the exact formulation of this second law?

Maxwell In fact there are several. The German physicist R. Clausius, one of the founders of thermodynamics, formulated the second law thus:

> **Second law, Clausius formulation:** [I]t is impossible to construct a device that, operating in a cycle, will produce no effect other than the extraction of heat from a cooler to a warmer body.[1]

Saint Peter I see. What are the others?

Maxwell A somewhat different formulation was given by Lord Kelvin, another influential researcher of thermal phenomena:

> **Second law, Kelvin formulation:** [I]t is impossible to construct a device which, operating in a cycle, will produce no effect other than the extraction of heat from a reservoir and the performance of an equivalent amount of work.[2]

[1] Sklar, 1993 [28], p. 21.
[2] Sklar, 1993 [28], p. 21.

As long as we restrict ourselves to systems whose absolute temperature is positive, these two formulations of the second law are equivalent. They tell us that we cannot simply take one hot object and use its heat to produce work: we always have to use two objects, one hotter that the other, and can only partially transform the heat of the hotter object to work. They tell us that there can be no machines which can take two vessels of gas of the same temperature and transfer heat from one to the other without changing the rest of the world.

Saint Peter I'm not at all an expert in physics, but I thought the second law asserted the non-decrease of entropy. Why doesn't the notion of entropy appear in the two formulations you've given?

Maxwell I'm coming to that. What Kelvin's and Clausius's versions of the law show us, is that we can define a state-function S (up to an additive constant):

$$\int_A^B \frac{đQ}{T} = S_B - S_A, \tag{1}$$

where A and B are two points in state-space, T is the absolute temperature, $đQ$ is the inexact differential of the heat Q and S is the entropy. The second law of thermodynamics tells us that if a system does not exchange heat with its surroundings – in other words, if it is adiabatically isolated – no processes can take place which lower its entropy S. Explicitly:

Second law, entropy formulation: Whenever an adiabatically isolated system evolves from equilibrium state A to equilibrium state B, the following relation holds for the entropy: $S_B - S_A \geq 0$.

The Clausius and Kelvin formulations of the second law explicitly forbid certain processes in which the entropy decreases to take place.

Saint Peter And these results hold for any system whatsoever?

Maxwell Yes. Perhaps surprisingly, they are sufficiently general to imply the non-occurrence of all such processes for any system you can think of. It is exactly this generality which made the second law one of the cornerstones of thermodynamics. But let's walk to my home while I continue the story. There is an old letter there which I'd like you to see.

Saint Peter Certainly.

Scene II: Atoms in the void

A quiet road in Heaven. Maxwell and Saint Peter walk amiably along it, now and then greeted by an angel or other inhabitant of the Celestial City.

Maxwell Time, however, leaves no things unchanged – not even fundamental physical laws. As the conception of matter as consisting of discrete atoms was increasingly accepted, physicists tried to replace thermodynamics – which ignored the structure of matter – with a science that would describe thermal phenomena in terms of interacting particles. Because of the immense amount of these contained in even the tiniest macroscopically visible systems, a full description of the positions and velocities of all the atoms was out of the question; instead, a statistical treatment was called for, calculating average values rather than individual properties.

Saint Peter So you created a new theory to deal with the same phenomena, only from a different perspective?

Maxwell Precisely. The new science which attempted to do this was called statistical mechanics, and one of the main aims of its founders – among whom I myself and especially Ludwig Boltzmann were prominent – was to reproduce the results of thermodynamics, including the second law.

Saint Peter Of course, because thermodynamics was so successful. Did you succeed in this enterprise?

Maxwell Not quite, I must admit – but this failure was a kind of success. Let me elaborate an example, which lies at the heart of the demon matter.

According to thermodynamics, it is not possible that two objects of the same temperature, when brought into contact with each other but isolated from the rest of the universe, will evolve in such a way that one becomes hotter while the other becomes colder. What would such a process look like in a statistical mechanical description? Both objects are thought to be composed of tiny particles which either vibrate about a fixed position (in solids) or can move around more or less freely (in liquids and gasses). The temperature of an object is proportional to the mean square velocity of its particles, in such a way that the greater the velocity of the particles that constitute an object, the greater its temperature.

Saint Peter Yes. The faster the molecules, the higher the temperature. I think Ludwig told me that once.

Maxwell What is very important for the rest of my story, and what I want you to understand very well, is that the atoms or molecules of a gas do not all have the same velocity. For gasses I have shown that the velocity distribution of the particles is the so-called Maxwell distribution:

$$f(v)dv = 4\pi(\frac{m}{2kT})^{3/2}v^2 exp(-mv^2/(2kT))dv. \qquad (2)$$

In this formula $f(v)dv$ is the chance that a certain particle has an absolute velocity between v and $v + dv$, m is the mass of the particles, k Boltzmann's constant and T the temperature. This

guarantees that although most particles have a velocity near to the average velocity, a fair amount of them has velocities notably higher or lower. Do you understand that?

Saint Peter Yes, I do. Some molecules are faster, some are slower, even within a gas with a uniform temperature. What is so relevant about that?

Maxwell That will become clear pretty soon, when I have finished my example. Please be patient.

If we put two containers, A and B, next to each other, both containing the same gas at the same temperature, and open a hole between them, what will happen is the following. Molecules will be able to pass freely through the hole, so some particles from A will move to B, and vice versa. Because the temperatures of the gasses are identical, so are the velocity distributions; on average, the molecules going from A to B will have the same square velocity as those going from B to A. As long as this happens the average square velocity, and hence the temperature, of the containers will not change. It is nonetheless *possible* that for a significant amount of time more fast molecules move from A to B than from B to A, and more slow molecules move from B to A than from A to B. In that case, the temperature of B will rise and that of A will fall, in a direct violation of the second law of thermodynamics. Statistical mechanics thus allows violations of the second law. Because of the huge amount of particles involved, the probability of such a fluctuation which gives rise to a significant temperature difference is extremely small. The situation is somewhat analogous to a box full of evenly mixed tiny red and blue balls. We can shake the box and thus throw about the balls in a random fashion, but chances are that this will not produce a nicely ordered configuration in which most red balls are in the left side of the box, and most blue balls are in the right side. The greater the number of balls, the less likely this spontaneous sorting is – and two containers of gas can easily hold 10^{23} 'balls'. In practice we can ignore the chance that any significant fluctuations in temperature will occur, because they are extremely unlikely. Although the second law does not hold unconditionally in statistical physics – violations *are* possible – it holds *probabilistically*: significant violations of the second law are, in general, highly unlikely.

Saint Peter Your example is very clear. I take it that when you spoke of your failure to derive the second law of thermodynamics actually being a kind if success, you were referring to this. By doing statistical physics, you discovered that the second law of thermodynamics is actually invalid.

Maxwell Precisely. And with that insight, we've reached my house. Please enter.

Scene III: Maxwell's Demon

Maxwell's study in the Celestial City. It is a spacious and light room, cluttered with books. Maxwell and Saint Peter occupy two arm-chairs in the centre, sipping a refreshingly cool white wine.

Maxwell I told you that significant violations of the second law are highly unlikely. This is true, but with one qualification: these violations are only unlikely as long as the demon known on earth as *Maxwell's Demon* is not around. My tiny friend, you see, has the ability to cool one gas and heat the other without changing anything in the environment, violating the second law whenever it wants and as badly as it wants.

Saint Peter So *that*'s his trick! But how does he do it, violating this time-honoured and – I might add – God-given law of nature?

Maxwell I first mentioned the demon in a letter to P.G. Tait in 1867. Maxwell starts rummaging through a pile of old letters next to his chair. Imagine two vessels of gas A and B, I said. Let A be hotter than B, and the two vessels separated by an isolating wall. In this wall is a small hole with a massless, frictionless slide in front of it which can be used to open and close the hole. Ah, here it is! He triumphantly shows a very old piece of paper, and hands it to Saint Peter. Please, read it aloud.

Saint Peter So this is the letter you spoke of. Let me see:[3]

> Now conceive a finite being who knows the paths and velocities of all molecules by simple inspection but who can do no work except open and close a hole in the diaphragm by means of a slide without mass.
>
> Let him first observe the molecules in A and when he sees one coming the square of whose velocity is less that the mean square velocity of the molecules in B let him open the hole and let it go into B. Next let him watch for a molecule of B, the square of whose velocity is greater than the mean square velocity in A, and when it comes to the hole let him draw the slide and let it go into A, keeping the slide shut for all other molecules.
>
> The number of molecules in A and B are the same as at first, but the energy in A is increased and that in B diminished, that is, the hot system has got hotter and the cold colder and yet no work has been done, only the intelligence of a very observant and neat-fingered being has been employed.
>
> Or in short if the heat is the motion of finite portions of matter and if we can apply tools to such portions

[3] Garber, Brush, Everitt, 1995 [14], p. 176-178. I have changed all abbreviations, such as 'vel.' for 'velocity', to their full counterparts for greater readability.

of matter so as to deal with them separately, then we can take advantage of the different motion of different proportions to restore a uniform hot system to unequal temperatures or to motions of large masses.

Only we can't, not being clever enough.

Maxwell And so it is. Such a simple scheme.

Saint Peter Your demon, then, is a being which can observe molecules individually, and can therefore sort them into faster and slower ones?

Maxwell Quite. The different motions of individual molecules, which are useless to us since we can only control them *en masse*, can be exploited by an observant and neat-fingered being because it *does* have the ability to control them individually. Because he has such neat fingers, my demon can control the atoms and molecules of matter and break the second law whenever he feels like it!

Saint Peter I can imagine he created quite a stir. And... well, there he is! A VERY SMALL, FRIENDLY-LOOKING DEMON JUMPS THROUGH THE WINDOW, WALKS UP TO MAXWELL AND SETTLES ON HIS LAP.

Maxwell Quite a stir indeed! But maybe you'd like to tell Peter about the enmity you encountered yourself.

Demon THE DEMON SPEAKS IN A VERY HIGH VOICE, BUT WITH AN IMPECCABLE OXFORD-ACCENT. Certainly, old chap. It will be my pleasure. But allow me to speak about the good things I've done first, before I turn to my critics.

Part I

The contingent second law

Chapter 1

The question of contingency

Throughout the centuries, physicists have formulated many 'laws of nature'. Among them are well-known ones such as Newton's laws of inertia and gravitational attraction, and Einstein's field equations of general relativity. The two main laws of thermodynamics – the first, which expresses the conservation of energy, and the second, which forbids the occurrence of entropy decrease – also belong to the exalted temple of famous physical laws. Yet the second law, which together with Maxwell's Demon is the protagonist of this thesis, is a very strange member of this company. This chapter will describe that strangeness, which centres around two features: the statistical nature of the second law, and the question whether it is a real law at all or merely a contingent generalisation. At the end of the chapter, Maxwell's Demon is introduced and the role it can play in solving the problems concerning the second law is discussed. It is argued that the demon cannot throw any light on the statistical nature of the second law, but can help us to answer questions of contingency and necessity.

1.1 The statistical law

1.1.1 A false law?

As indicated in the prologue, the second law is – in its thermodynamical formulation – simply false. In the Clausius formulation, the law is expressed thus:

Second law, Clausius formulation: [I]t is impossible to construct a device that, operating in a cycle, will produce no effect other than the extraction of heat from a cooler to a warmer body.[1]

The second law thus forbids absolutely the transport of heat from colder to hotter bodies by any device 'operating in a cycle'. The accuracy of this law was not widely questioned until the atomic hypothesis of matter became widely embraced by the physical community. When the theory of statistical physics was developed by Maxwell, Boltzmann and others, it became clear very quickly that the second law of thermodynamics could not hold unconditionally. I will illustrate this with a thought experiment.

[1] Sklar, 1993 [28], p. 21.

Suppose one has two containers, A and B, filled with the same gas at the same pressure, A hotter than B, completely isolated from their surroundings except for the fact that there is a tiny hole in the wall connecting them. Molecules can move freely from one container to the other through the hole, but it is so small that there is never more than one molecule making this transition at the same time. In statistical physics, the thermodynamical, macroscopic quantity *temperature* is equated – up to a constant of proportionality – with the 'statistical', microscopical quantity *average kinetic energy of the particles*. Hence, the average kinetic energy of the particles in A, $\frac{m}{2}\overline{v_A^2}$, where m is the mass of the particles and $\overline{v_A^2}$ is the average square speed, is higher than that of the particles in B, $\frac{m}{2}\overline{v_B^2}$.

Let one particle with velocity v_1 move through the hole from A to B, and another with velocity v_2 from B to A. It is clear that the average kinetic energy of A is decreased and that of B is increased if and only if $v_1 > v_2$, whereas the opposite effect takes place if and only if $v_1 < v_2$. In the first instance, heat is transported from a hotter to a colder body, in accord with the thermodynamical second law. In the second, heat is transported from a colder to a hotter body, in violation of the thermodynamical second law. Now the speeds of the molecules of an ideal gas at temperature T are described by the Maxwell-velocity distribution

$$f(v)dv = 4\pi(\frac{m}{2kT})^{3/2}v^2 exp(-mv^2/(2kT))dv. \tag{1.1}$$

This is a distribution with tails extending to zero and infinity, so there will be some molecules in B which are faster than some molecules in A, even though A is at a (much) higher temperature than B. If there are molecules in B which are faster than some molecules in A, the possibility that $v_1 < v_2$ in any process of particle exchange is non-zero. The second process, then, in which heat is transported from a colder to a hotter body, is not forbidden by statistical physics. It can happen that – by sheer coincidence – an exceptionally fast molecule from the colder gas is exchanged for an exceptionally slow molecule from the hotter gas. This process is allowed and actually takes place, but it contradicts the Clausius formulation of the second law. Thus, the second law of thermodynamics is literally false. Literally – but there seems to be large grain of truth in it. What is this grain, and how can the second law be reformulated in statistical physics, so as to make it true once more?

1.1.2 The statistical formulation

Recapitulating: in thermodynamics, that claim was made that decrease of entropy (or any of the effects associated with it, such as those forbidden by the formulations of Clausius and Kelvin) is strictly forbidden and never occurs. The development of statistical mechanics showed that this unconditional second law does not hold; there is always a possibility, extraordinarily small though it may be, that entropy will decrease. One might hope that the possibility for any downward entropy fluctuation is so small that we will never be able to observe one. Then, one might be tempted to adopt the following reformulation of the law: although entropy decreases are possible, it is overwhelmingly likely for any macroscopic system that no detectable violation of the second law takes place. Unfortunately, these hopes are not fulfilled, and the proposed reformulation is not successful: some macroscopic fluctuation phenomena, such as Brownian

motion, are easily detectable, even though they constitute violations of the thermodynamical second law. We need to search for another way of improving the second law.

Let us reconsider the thought experiment of the last subsection. If A is at a much higher temperature than B and molecules for exchange are chosen at random, the chance of the molecule leaving A being faster than that leaving B is overwhelming; this can easily be seen from the Maxwell-distribution 1.1. So chances are that the great majority of all particle exchanges will result in a heat flow from the hotter to the colder body, and the opposite will happen only relatively rarely. Hence, although many small downward fluctuations in entropy take place, they are generally more than compensated for by a much larger amount of interactions in which the entropy increases. And no processes exist, as far as we know, which ensure that only the entropy decreasing exchanges take place; no machines can be built, it seems, which create a situation where the unlikely exchanges which lower the entropy become dominant and create with high probability a major entropy decrease. Through these considerations, Smoluchowski (1912, [30]) realised that although in almost any case decrease of entropy is a possibility and may actually be observed, the chance of such decreases going on for a significant time is very slim. Following his lead, we might reformulate the second law thus:

Statistical second law, entropy formulation: Entropy is not forbidden to decrease, but in all processes the probability of continuous and macroscopically significant entropy decrease is extremely small.

What is unfortunate about this formulation of the statistical second law is that the concept of entropy is not uniquely defined in statistical physics: instead, there are a number of different and competing definitions of entropy, for some of which the above formulation is simply false. These formulations and their relationship with the second law will be discussed at length in chapter 2. Presently, it is more useful to describe the statistical second law in terms of the processes we wish to forbid: those which take random molecular motion (heat) and completely convert it to macroscopically useable work. Such a 'Kelvin' version of the second law was actually proposed in Smoluchowski 1912.

Statistical second law, 'Kelvin' formulation: It is impossible to construct a device which with a high probability, operating in a cycle, will produce no effect other than the extraction of a macroscopically significant amount of heat from a reservoir and the performance of an equivalent amount of work.

This formulation is, barring negative absolute temperatures, equivalent to a 'Clausius' formulation which forbids a reliable device operating in a cycle to do nothing else but transport heat from a colder to a hotter object. Neither of the two, however, is equivalent with the entropy formulation of the statistical second law, if only because the latter is ambiguous. In fact, I will later claim that they are not equivalent for any of the non-ambiguous notions of entropy that are generally used either. I suggest that the basic idea behind the second law of statistical physics is best captured by the 'Kelvin' or 'Clausius' formulation. Whenever I speak of 'the statistical second law' in the rest of the text I will mean the (phenomenal) Kelvin/Clausius formulation, unless otherwise specified. I will

often simply call it 'the second law' if there is no need to specifically distinguish the statistical from the unconditional thermodynamical version.

Let us take a closer look at the statistical second law. It contains one very strange clause, one thing that distinguishes it from all the other laws of physics. The second law speaks about devices which 'with a high probability' fail to do something; it forbids certain things from happening very often, but allows that they may happen occasionally. Every other law of classical physics,[2] be it Newton's laws of motion, Einstein's field equation, Maxwell's theory of electromagnetism or anything else, forbids only absolutely. Newton's laws tell us that a body on which no forces work undergoes no acceleration; it does not tell us that the body has 'a very big chance' of undergoing no acceleration. Maxwell's theory tells us that two positive charges far removed from any other charge distributions will repel each other; not that they will repel each other 'with high probability'. Somehow or other, then, probability and chance make an appearance in the second law of statistical physics, although they make no such appearance in the rest of classical physics. This is the first strange feature of the second law.

1.2 Reduction and contingency

1.2.1 Reduction

We now turn to the other strange feature of the second law. It has been observed that the law does not describe new physical interactions. Indeed, it says nothing at all about microscopic processes. If I have a container of gas, the second law tells me nothing about the interaction between the gas molecules, about the ways in which they will collide or about the collisions between the molecules and the walls of the container. The microscopic laws of motion are given to us by mechanics and electrodynamics, and they are completely determined by these theories.[3] Statistical physics does not add one iota to our knowledge in this respect. But if we know the detailed behaviour of every single particle of the gas – what is there left to know? Is there still some independent feature of the gas left undescribed which could be the subject matter of the second law? Surely not. If we know the position and velocity of every molecule in the gas and of every molecule of the container at every moment in time, then we know everything which can be possibly known. We know whether temperature differences are created or destroyed; we know whether heat is changed into work. If we know the laws of microscopic behaviour, we seem to know everything.

But this puts the second law in a strange position. Anything it tells us about the behaviour of systems must in some way be reducible to statements about the microscopic evolution of these systems; and hence, the claims of the second law must be reducible to those of the laws of microscopic behaviour. At least

[2]This thesis is not the place to discuss the probability-related aspects of quantum mechanics. It ought to be remarked, however, that the probabilistic aspect of statistical physics has nothing to do with quantum effects.

[3]Of course, we might actually need general relativity, quantum field theory and whatnot to describe the gas with complete accuracy; and maybe not even these would be enough. But just assume, for the sake of argument, that we have a complete and accurate theory of the microscopic behaviour of the gas. It is clear in any case that the second law is *irrelevant* to this description.

at first glance, the statistical second law must be reducible to more basic, more fundamental laws of nature.

There is at least one well known problems with this reduction. In the formulation we gave of the second law, machines are forbidden which with high probability continuously create work from heat. There is no similar prohibition of a machine which changes work into heat – indeed, work changes to heat every time there is friction. The second law is therefore time asymmetric: a process wherein heat is converted to work is forbidden, but the time-reverse, a process wherein work is converted to heat, is allowed. This becomes a problem for attempts to reduce the second law to laws of microscopic motion, as all laws of the latter type known to us are completely time-reversal invariant. And surely a time-asymmetric law cannot be proven from completely time-symmetric assumptions. The problem of squaring the time-asymmetry of the statistical second law with the time-symmetry of the underlying dynamics has generated a huge amount of literature – see, for instance, Reichenbach 1956 ([25]), Horwich 1987 ([16]), Sklar 1993 ([28]), Albert 2000 ([1]).

We will not enter into this debate. Questions of time asymmetry – Is the second law really time-asymmetric? How does this relate to its possible reduction to more fundamental laws? Has entropy anything to do with the 'arrow of time'? – are not our concern in this thesis, and discussing them in any detail will be studiously avoided. But this topic can nevertheless serve as a useful illustration of the notions of necessity and contingency to which we presently come.

1.2.2 Necessity and contingency

We may wonder what the status is of the statistical second law, *given* the validity of the laws of microscopic motion. In the remainder of this part of the thesis I will focus exclusively on models from classical mechanics, so I'll likewise restrict the discussion to the classical mechanical laws of motion. Thus, I'll ignore electrodynamics and other theories which in reality are very important in describing the paths and interactions of particles in a gas. My main reason for focussing on classical mechanics is that it furnishes us with a relatively simple case in which all the major issues I want to talk about can already be discussed in the necessary detail. Adding electrodynamics or quantum mechanics would greatly increase the difficulty of most discussions, without yielding any real benefit in terms of clarity and understanding.

We may, I repeat within the new context just defined, wonder what the status of the statistical second law is, given the validity of the laws of classical mechanics and their completeness concerning the detailed description of every system. Do the laws of classical mechanics necessitate the validity of the second law, or is the second law's validity contingent on still further facts about the world? The dichotomy can be spelled out as follows. Either the second law is *necessary*, or it is *contingent*. We'll say that the second law is *necessary* if it follows from the universal validity of the laws of classical mechanics,[4] supplemented if need be by some elementary considerations of probability, of ensembles and suchlike. We'll say that the second law is *contingent* if it does

[4]Remember that I'm restricting myself, in this first part of the thesis, to models of classical mechanics.

not follow from these things alone, but further facts about the world are needed to derive it; if such facts exist, the second law is *contingent on* them. Thus, if the second law only holds because there are no tiny intelligent creatures in our world, but the existence of such creatures is merely a brute fact and not necessitated by the laws of classical mechanics, then the second law is contingent. It is contingent on the non-existence of tiny intelligent creatures. If, on the other hand, such creatures are indeed the only possibility of breaking the law but they are impossible given classical mechanics, then classical mechanics alone is enough to save the day for the second law. In that case, it is necessary.

By way of further illustration, let me return to the subject mentioned in the previous subsection. The dichotomy between necessity and contingency plays a central role in the discussions about time-asymmetry. Many people, for instance Paul Horwich (1987, [16]) and David Albert (2000, [1]), invoke a kind of contingency to solve this problem. According to them, the initial state of the universe was a very special state, with few correlations (Horwich) or very low entropy (Albert). The only reason that we currently have entropy rising wherever we look, instead of falling or being more or less constant, is because of the special nature of the initial state of the Universe after the Big Bang. According to them the Second Law is *contingent on* the occurrence of this special state. Whether their reasoning is correct is a delicate and complex question, and we will not venture onto this battle-scarred terrain.

1.2.3 Is the second law contingent?

Wondering about the contingency or necessity of the fundamental laws of nature is something best left to theologians and metaphysicians of a speculative kind. But in the case of the statistical second law, which for all the world does not look like a fundamental law of nature, the question is a valid one for philosophers of physics. Even leaving aside the questions of time asymmetry, it is not obvious whether the second law can be derived from the fundamental laws of physics. In our case, we want to know whether it can be derived from classical mechanics. Is the statistical second law valid in all models of classical mechanics, or is there a large class of models in which the law is simply false? This question is very tricky. We have to be very precise about the criteria we use to determine whether something is a 'large class of models'; we have to spin some subtle arguments about the introduction of ensembles and probabilities; and we have to determine when the second law holds or fails to hold in a model of classical mechanics – which is not trivial, as neither 'entropy' nor 'heat' are terms which we customarily use in this theory.

This first part of the thesis is concerned with exactly this question about the contingency of the second law. It will try to be precise, talk about ensembles, present models of classical mechanics, formulate a notion of the validity of the second law – and in the end, I will claim that the statistical second law is contingent, not necessary. I propose that this discussion is at the heart of the Maxwell's Demon problem – although other authors have not, of course, shared my self-imposed limitation to classical mechanics –, that the demon's main purpose is exactly to help us solve the question of necessity and contingency. I will try to support that claim in the next section.

1.3 What we learn from the demon

The demon is a thought experiment: Maxwell never believed that such beings existed, or that men could make tiny machines that functioned like them. This negative attitude towards the possibility of demons has been dominant in the literature ever since, but it has been exemplified in two very different ways. Some people saw the demon as a threat to an accepted and successful part of physics – the statistical version of the second law of thermodynamics – and tried to prove from physical principles that it could not exist. Such attempts will be discussed in part II. Others, like Maxwell himself, accepted the non-existence of demons as obvious or unproblematic, and wished to make use of them as explanatory or pedagogical devices. But just *what* they can explain deserves some clarification.

In the preceding sections, I discussed two strange features of the second law: its statistical nature, and its possible contingency. Maxwell's Demon can be interpreted as a helpful thought experiment for understanding either of these two features. I will presently attack the idea that the demo can teach us anything about the second law's statistical nature, and defend the view that it is a teacher on contingency and necessity.

1.3.1 'Only statistical certainty'

In an undated letter to Peter Guthrie Tait, Maxwell reflects on the history of his demon in a few short sentences[5]:

1. Who gave them this name? Thomson.[6]
2. What were they by nature? Very small **BUT** lively beings incapable of doing work but able to open and shut valves which move without friction or inertia.
3. What was their chief end? To show that the 2nd Law of Thermodynamics has only statistical certainty.

In this letter Maxwell suggests that the main point of his demonic thought experiment is that it enables us to see clearly that the second law has only statistical certainty. But what does he mean when he makes this claim? 'Statistical certainty' is an obscure notion, because its most obvious interpretation has changed since Maxwell's time. With the 'statistical approach' to physics, Maxwell meant the practice of not asking questions about individual molecules, but only about gross macroscopic averages. That the second law has only statistical certainty would mean, for him, that the second law is only true if we limit ourselves to these averages and 'avoid all personal enquiries of molecules'. In his letter to Tait, Maxwell restates in different words his original claim[7] about the demon: that it shows us that the second law does not hold for tiny-neat

[5] "Catechism on Demons", Garber, Brush and Everitt, 1995 [14], p. 180.

[6] Maxwell conceived of a 'very observant and neat-fingered being'. It was William Thomson, also known as Lord Kelvin, who dubbed this creature a 'demon'.

[7] See page 12: *"Or in short if the heat is the motion of finite portions of matter and if we can apply tools to such portions of matter so as to deal with them separately, then we can take advantage of the different motion of different proportions to restore a uniform hot system to unequal temperatures or to motions of large masses. Only we can't, not being clever enough."* – Garber, Brush, Everitt, 1995 [14], p. 177.

fingered beings, but only for clumsy people like us who are not clever enough – that is, who must use statistical methods.[8]

But today 'statistical certainty' means something different: a claim that the second law has only statistical certainty will be interpreted by almost anyone as a claim that it does not hold always, but only with a high probability. 'Statistical' no longer points to a method, but to a use of probabilities. For many modern readers, it will appear to be obvious that Maxwell wishes to claim that the thermodynamic version of the second law does not hold; only its statistical version does. It is therefore of the highest importance to carefully distinguish between these two interpretations of Maxwell's words: the one in which it is advocated that the second law holds only because we are limited to considering and manipulating atoms *en masse* and the one in which it is claimed that the second law only holds with high probability, not unconditionally. I will now argue that using the second interpretation, Maxwell's claim about the demon is undefendable.

We can draw a distinction between three possible cases: the second law holds in its thermodynamic version, unconditionally; the thermodynamical version does not hold, but the statistical second law does; or not even the statistical second law holds. Maxwell's almost infinitely repeated letter can be construed as claiming that the demon clarifies the distinction between the first and the second possibility, showing that the statistical version of the law has to be preferred to the unconditional one. Yet the demon is a creature which can break the second law, *even* its statistical version. Wherever the demon is present, neither of the first two options is correct and we would have to opt for the third.

In fact, the distinction between the first and second option can be made clear if we just imagine two containers of gas, one hot and one cold, connected by a small hole – but without a shutter or a demon. It is easy to see, once we have knowledge of the Maxwell velocity distribution, that there is a non-zero chance of fast molecules moving from the cold to the hot gas, and slow molecules moving the other way; complemented with an explicit calculation of the probabilities involved, this is all that is needed to show that the second law does not hold, but only its statistical version does.[9] Adding a shutter and a demon does not just complicate the example unnecessarily, it positively destroys its ability to show the statistical validity of the second law. If a demon can exist – and it is not immediately clear that it cannot – the statistical second law is false. Maxwell's Demon is unfit to show that 'the 2nd Law of Thermodynamics has only statistical certainty', *if* 'statistical certainty' is interpreted in the modern way rather than the way Maxwell meant it to be understood.

Confronted with the two strange features of the second law which I identified, its statistical nature (in the modern sense of the word) and its contingency, the demon certainly cannot clarify the former. *Ex hypothesi* Maxwell's Demon *breaks* the second law in both its thermodynamic and its statistical form; it does not tell us why or in what sense the second law has a statistical nature. (Although an understanding of this aspect of the second law might be very important for accurately discussing the demon – as a necessary preliminary of

[8] See for instance Heimann 1970 ([15], especially p. 62-67) for a discussion of Maxwell's thoughts on the 'statistical method'.

[9] At least, such a calculation would show that the statistical second law is the strongest law that *can* hold, and that the example does not give us any reason to doubt that it *does* hold. It does not rigorously prove that the statistical second law is true.

the discussion, not as a result of it.)

Unfortunately, Maxwell's claim has been quoted very often without an appropriate discussion of its meaning, leaving the unsuspecting reader in great danger to misinterpret it. Two recent and representative examples are Leff and Rex 1990 ([19], p. 5; also the 2003 renewed version, [20], p. 5) and Bub 2002 ([9], p. 2). Leff and Rex quote Maxwell's letter to Tait, which I cited above, then fail to explain his notion of 'statistical certainty'. This alone could all too easily leave the reader with the wrong idea of Maxwell's intentions. But in addition, they claim in the preface of their book (p. vii) that the demon can teach us about 'the role of probability and statistics'. This is precisely what it can not do. Describing the birth of the demon, Leff and Rex claim (p. 4): 'Maxwell's thought experiment dramatizes the fact that the second law is a statistical principle that holds almost all the time for a system composed of many molecules'. It does not dramatise the fact that it holds almost all the time; instead, it points the way to situations in which it does not hold at all. Jeffrey Bub tells us that the point of the demon argument was to show that the second law has 'only statistical certainty', without explaining that notion. But in the next sentence, he speaks about 'statistical fluctuations', which reveals a use of the modern meaning of 'statistical'. Given this widespread lack of attention to and confusion about the proper meaning of Maxwell's claim, and the untenability of its modern interpretation, I think it has been valuable to point out at some length that the demon can *not* be used to show that only the statistical second law holds.

1.3.2 Dissection: contingency and the demon

The proper purpose of the demon is to tell us about the necessity and contingency of the second law. It is a thought experiment which can be used to decide the question whether the law is contingent, and if it is, the demon can tell us what the second law is contingent *on*. This point bears some explanation.

Suppose that all demons we can think of, all creatures and machines which change heat into work seemingly effortlessly, violate the laws of classical mechanics. (Such demons are actually easy to construct – once we allow arbitrary non-Hamiltonian force fields, the second law can be broken at will. See for instance Zhang and Zhang 1992, [33].) In that case, the validity of the laws of classical mechanics forbids the existence of demons, and hence necessitates the validity of the statistical second law. The second law would be necessary. If, on the other hand, we can construct a demon which is a model of classical mechanics, the second law does not hold of necessity. Supposedly, this demon will not exist in our world for some reason or another – I am assuming the statistical second law may turn out to be contingent, but will not turn out to be false –, some reason which is not implied by the laws of classical mechanics, such as a very special initial state of the universe. We have then proven that the second law is contingent on this reason. Hence constructing demons, or disproving the constructibility of certain classes of demons, answers our questions about the necessity and contingency of the second law.

In other words: Maxwell's Demon can violate the second law by definition, but we do not believe that the second law is violated in our world. Thus, those features of the demon which are not exemplified by any real objects will be the features that enable it to violate the second law. The validity of that law is

contingent on the non-existence of those features in our world. In other words, by dissecting a demon, by exploring the traits that enable it to violate the law, we may discover *why* the second law holds: examining a demon is a case-study in the 'transcendental conditions of the validity of the second law'. Possibly there is a large variety of demons around, which can break the second law because of very different attributes; in that case, the validity of the second law holds because of the non-occurrence in reality of *all* such attributes. What Maxwell's Demons can give us, when properly dissected, is potentially deep insight into the foundations of statistical physics; in particular, they may provide us with answers to the question: which properties of reality ensure the validity of the statistical second law? Are these properties fundamental laws (in which case the second law is necessary), or are they simply facts (in which case the second law is contingent)?

I suggest that it was to provide answers to questions like these that Maxwell's Demon was first summoned. Maxwell wrote, in the letter quoted at length on page 12:

> [I]f the heat is the motion of finite portions of matter and if we can apply tools to such portions of matter so as to deal with them separately, then we can take advantage of the different motion of different proportions to restore a uniform hot system to unequal temperatures or to motions of large masses. Only we can't, not being clever enough.

The suggestion is that *if* we could only apply very tiny and well-made tools to individual atoms, *then* we could break the statistical second law at will. Hence the validity of this law rests upon the fact that we cannot make such tools, that, in other words, we are not 'clever enough'. What it means to be 'clever' is of course a matter of debate. But I think it is not far-fetched to construe Maxwell's letter as saying that the validity of the second law is contingent on the non-existence of creatures with the ability to efficiently apply tiny tools to molecules in motion. How this claim can be made precise and whether Maxwell was right are the topics which will occupy us from now on.

Chapter 2

Portrait of the Demon

Our quest is to understand the necessity or contingency of the second law by looking at beings which can break it and examining their features. These beings are called Maxwell's Demons. Before we can investigate whether or not a Maxwell's Demon can be constructed in classical mechanics, we need to know when a creature deserves this ominous name. We need to be able to sift the pretenders from the *bona fide* demons. We need to sketch a portrait of the Demon.

First of all, this means that we have to find a criterion for breaking the second law. This is not as easy as it sounds, for merely saying that the demon ought to 'decrease entropy' is too vague. Within statistical physics, there are several notions of entropy, and not all of them are connected to the second law in a clear way. The first aim of this chapter will therefore be to review the possible criteria for violating the second law, making use of phenomenal considerations, Boltzmann entropy, fine-grained Gibbs entropy and coarse-grained Gibbs-entropy. I will decide in favour of the phenomenal criterion, as it is the only one connected to the second law in an obvious way. That this choice is very significant will become clear in chapter 4.

Secondly, the demon ought to be able to change heat into work when confronted with any of a large class of systems. It is perhaps quite trivial to think up a system which can act as a Maxwell's Demon for one particular configuration of a gas (a carefully timed trapdoor might do the trick – though some reservations can be made), but we want our demon to be able of handle all kinds of initial configurations. We will try and find a good formulation of this requirement in section 2.4.

Thirdly, we have to take a look at the notion of cyclicity. The demon ought to be able to change heat into work, 'operating in a cycle' – it is, for instance, not good enough if the demon uses a battery to change heat into work, expending lots of irrecoverable potential energy to make the change from heat into work. But in what sense must the demon operate in a cycle? This question is discussed in section 2.5.

These three points complete the portrait of the demon as it will be sketched in this chapter. It will give us all the tools we need to get on with the real work: constructing or disproving demons.

2.1 Breaking the law: the phenomenal approach

2.1.1 Phenomena and entropy

Maxwell's Demon is intimately connected to the statistical second law, as it is by definition a being which can break this law. It is therefore of crucial importance to understand when this law is broken, if we wish to be able to recognise demons when we see them.

The thermodynamic second law claims that entropy cannot decrease, but always has to increase. This suggests an equivalent formulation of the statistical second law in terms of entropy: 'no machine exists which can with high probability lower the entropy of a system for a significant amount of time', or something like that. Unfortunately, there is no single universally accepted definition of entropy in statistical physics. The Boltzmann entropy, the fine-grained Gibbs entropy and the coarse-grained Gibbs entropy are all used more or less frequently, yet they are very different from each other both in letter and spirit. The conclusions we will reach about the demon, and indeed about the contingency or necessity of the second law, are very much dependent on the criteria used to determine whether a potential demon does or does not break the second law. It is therefore of prime importance to specify those criteria in advance, or at least be as clear about them as possible. Presumably, as physicists we are not really interested in the decrease or non-decrease of a mathematically defined quantity, unless that quantity tells us something worth knowing about the physical world. Hence, we are only interested in the behaviour of the three competing quantities called entropy in so far as they inform us about the success or failure of Maxwell's Demon to do what it ought to do: produce large-scale anti-entropic effects such as the demon in Maxwell's original thought experiment did. It is a good idea to specify what kind of effects would constitute anti-entropic behaviour, and then subject the different definitions of entropy to a critical test: can they provide us with a reliable indication of such behaviour? So we will first establish a phenomenal criterion of violating the second law, and then look at the possibility of supplanting it with a more mathematical and formal criterion.

2.1.2 Phenomenology of demons

A violation of the second law of thermodynamics is easily seen to be possible from within the framework of statistical physics. The law only holds 'statistically', which we suppose to mean that there are no systems, operating in a cycle, which reliably produce large scale violations of the second law of thermodynamics. We start by repeating the Clausius and Kelvin formulations of the thermodynamical law.

Second law, Clausius formulation: It is impossible to construct a device that, operating in a cycle, will produce no effect other than the extraction of heat from a cooler to a warmer body.

Second law, Kelvin formulation: It is impossible to construct a device that, operating in a cycle, will produce no effect other than the extraction of heat from a reservoir and the performance of an equivalent amount of work.

The two phenomena mentioned in these laws, when produced reliably and on a large scale by a system in a cycle, would surely convince us that the system in question is a Maxwell's Demon. It should be noted that, using Carnot engines, a temperature difference can be exploited to convert heat to work, and work can be used to heat and cool objects. Therefore, any demon that produces one of the two phenomena will, with some simple tools, also be able to produce the other. Another useful example of an anti-entropic effect is the creation of a pressure difference between two containers of gas which were initially at the same pressure and temperature. This can be done by a somewhat less intelligent sibling of Maxwell's original demon: if it only opens the slide to let molecules pass from A to B regardless of their velocity, but never allows molecules to pass the other way, the pressure in A will drop while that in B will rise. This pressure difference can be exploited by a little turbine, which will change the gas's kinetic energy (the heat) into work. A 'pressure demon' can, with some simple tools, convert heat into work; as this can also be done the other way around, a reliable producer of any of the three phenomena mentioned can be easily turned into a system that can produce all of them at will. Phenomenally, then, the following is a good characterisation of Maxwell's Demons:

Maxwell's Demon, phenomenal definition: A sufficient condition for a system to be a successful Maxwell's Demon is that, operating in a cycle, it can produce, with high probability, at least one of the following phenomena without making any other changes in the environment:

- Two systems at the same temperature evolve to one system at a significantly higher and one at a significantly lower temperature.
- A significant amount of heat is converted completely into work.
- Two vessels containing the same gas at the same temperature and pressure evolve to one vessel with a significantly higher and one vessel with a significantly lower pressure, while the temperatures remain equal.

This partial definition[1] captures the basic idea behind the demon very nicely, except for the vague clause about cyclicity, which is in need of further elucidation, and the non-specification of the range of systems on which it has to operate successfully. We will return to these points in sections 2.5 and 2.4 respectively.

The main advantage of the phenomenal definition of Maxwell's Demon is that it clearly captures the physically relevant processes. If the phenomenal Maxwell's Demon exists, we have all the wondrous machines at our disposal that can furnish us with nigh unending amounts of useful energy. The main disadvantage of the phenomenal definition is that it lacks the mathematical clarity of more formal definitions. We will now try and find out whether any of the mathematical notions of entropy used in statistical physics can adequately capture the same physical effects as the phenomenal definitions.

[1] It is a partial definition, since it contains sufficient but no necessary conditions. For instance, a system which can sort two mixed gasses while keeping both pressure and temperature uniform, should also be seen as a Maxwell's Demon. I assume that any demon can be transformed into one of the above when it is equipped with some 'simple tools', but this last notion is too vague to allow for strict definition.

2.2 Breaking the law: Boltzmann entropy

2.2.1 Boltzmann entropy

The Boltzmann-entropy is explicitly constructed to reproduce the familiar thermodynamical fact that entropy almost always increases and almost never decreases.[2] The basic idea behind it is that although the number of (microscopic) states a macroscopic system can be in is enormous, most of these states cannot be observationally distinguished. There is no way we can measure the position and velocity of every molecule in a gas. Instead, our measurements give us access to only a small amount of information about the system: we are, say, able to measure its temperature and pressure as well as large internal fluctuations of these variables. Assuming that our measurement resolutions are always finite and the possible values of all our variables have both upper and lower bounds, there is a finite number of possible outcomes of any set of simultaneous measurements on a system. These sets of outcomes represent all we can know about the system, and are said to define *macrostates*: every possible combination of measurement results defines one macrostate. Any macrostate will correspond to a large amount of microstates, namely all those microstates which will yield the measurement results defining the macrostate. However, not every macrostate corresponds with as many microstates as every other; in fact, some of them will correspond with a vastly larger amount than others. Now imagine that the system is wandering more or less aimlessly[3] through the space of all its possible microstates; we will then expect it to spend much larger amounts of time in 'big' macrostates than in 'small' macrostates. If it starts in a small macrostate, chances are that it will evolve towards a big macrostate; and the second law tells us that if a system starts in a state of low entropy, chances are that it will evolve towards a state of high entropy. This analogy is the rationale for defining the Boltzmann-entropy of a microstate as proportional to the logarithm of the size of the macrostate it belongs to. The system will probably evolve from small to big macrostates; hence, from low to high entropy.

We will shortly turn to a mathematical definition of the Boltzmann-entropy, but now two important caveats need to be made. First, that deriving the statistical second law in the manner outlined above is not simple; in fact, it is one of the most complicated problems in the foundations of statistical mechanics and quite a lot of people (though by no means all) believe the effort to be fundamentally misguided. Second, that my reference to 'amounts of microstates' was misleading: because systems in statistical mechanics are mostly modeled as having a continuous state space, every macrostate will correspond to a non-denumerable infinity of microstates. To meaningfully speak about its size, we need to define a measure μ on the state space, which is not trivial. In the mathematical treatment below, we will assume that a measure has been chosen, but the reader is asked to keep in mind that this is not unproblematic.

Consider a system T with a state space Γ and a normalised measure μ defined on this state space.[4] At time t it has a microstate $\vec{x}(t)$ in Γ. A set of

[2] Whether it can achieve this aim is a very interesting and surprisingly hard problem.

[3] A notion which is, of course, unacceptably vague for any serious treatment of the question whether real systems 'wander aimlessly'. One way to make it more precise is to introduce the notion of ergodicity, but a discussion of that concept would carry us too far from this thesis' main subject.

[4] This mathematical discussion, as well as part of that of the other definitions of entropy,

macrostates $\{M\}$ is defined on Γ in such a way that every microstate belongs to exactly one macrostate; in other words, the macrostates form a partition of the state space. If $\vec{x}$ is a microstate, then $M(\vec{x})$ is the macrostate corresponding to that microstate; more precisely, $\forall\vec{x} : \exists M(\vec{x}) : \vec{x} \in M(\vec{x})$, $M(\vec{x}) \in \{M\}$. The measure of a macrostate $M(\vec{x}) \in \{M\}$ is $\mu(M(\vec{x}))$. There is, therefore, a map $\vec{x} \to M(\vec{x}) \to \mu(M(\vec{x}))$ from Γ to $\mathbb{R}^+$. The **Boltzmann entropy** is defined as:

$$S_B(\vec{x}) = k_B \ln[\mu(M(\vec{x}))], \tag{2.1}$$

where k_B is Boltzmann's constant. Since the logarithm of x is a monotonically increasing function of x, the Boltzmann entropy of a system increases when it evolves to a macrostate of greater measure and decreases when it evolves to a macrostate of lesser measure.

2.2.2 Boltzmann and phenomena

What would a demon have to do if it wanted to lower the Boltzmann entropy of a system? It would have to make the system evolve to a microstate associated with a macrostate that is significantly smaller than the macrostate in which the system started. And it would have to do this without creating a corresponding increase in the Boltzmann entropy of itself or the rest of the world. Since Boltzmann entropy is additive (the entropy of two systems together is the sum of their respective entropies), this implies that the Boltzmann entropy of the world[5] must go down as the demon operates. Thus, the demon must ensure that the world evolves from a macrostate with a large measure to a macrostate with a lesser measure, and it has to succeed in doing this with high probability. I will call the demons that can do this Boltzmann's Demons.

Now what is the connection between Boltzmann entropy and the phenomena listed in our phenomenal definition of the demon? It is not immediately clear that measure in state space has anything to do with heat and work, pressure differences and temperature differences; but such connections *do* exist, according to those who advocate the use of Boltzmann entropy. As an easy example, let us look at the pressure difference between two equally big containers, A and B, which contain a total of N particles of an ideal gas. Furthermore, assume that both containers are and remain at equal temperatures. The pressure in a container increases monotonically with the number of particles in the container, as we can see from the ideal gas law $P = N(k_BT/V)$. In the initial situation there are $N/2$ particles in A, and the same amount in B; obviously, they are then at equal pressure. We now wish to show that creating a pressure difference (something a Maxwell's Demon might do) constitutes the lowering of the system's Boltzmann entropy (something a Boltzmann's Demon might do), and *vice versa*.

To do this we first need to choose a partition of and a measure on the state space. To simplify matters, we use a discrete state space $\Gamma = [\mathcal{A}, \mathcal{B}]^N$, which is the set containing all sequences of $\mathcal{A}$'s and $\mathcal{B}$'s of length N. Assuming the N particles in the gas to be numbered, an $\mathcal{A}$ on the i-th place of such a sequence means that particle i is in container A, and the meaning of a $\mathcal{B}$ is analogous. Every possible distribution of particles over the two containers

is loosely based on Lavis, 2003 [18].

[5] Well, that part of it which is in causal contact with the operating demon.

A and B corresponds to one such sequence, and therefore to one state. The measure $\mu(X)$ of a set $X \subset \Gamma$ is the number of sequences in X divided by 2^N.

We still have to identify the macrostates: how do we choose a partition on this state space? The one observable we are interested in in this example is pressure, and the pressure in A and B is only dependent on the number of particles in those containers. Since the total number of particles is constant, there is only one independent measurable variable: the pressure of A, or, equivalently, the number of particles in A. We therefore define a set of macrostates $\{M\} = M_0, M_1, \ldots, M_N$, where $x \in M_i$ if and only if the sequence x contains exactly i $\mathcal{A}$'s. Elementary combinatorics gives us that $\mu(M_i) = N!/(i!(N-i)!)$, which attains a maximum for $i = N/2$ and monotonically decreases as the number of $\mathcal{A}$'s (or $\mathcal{B}$'s) differs more from this equilibrium value. Hence, the Boltzmann entropy of the system attains its maximum at equal pressures, and becomes lower the more the pressures diverge. A Maxwell's Demon that creates a pressure difference also lowers the Boltzmann entropy; and in the simple model we are currently considering decreasing the Boltzmann entropy also implies creating a pressure difference. The Boltzmann entropy functions exactly as we would like it to.

However, this example does not prove that decrease of Boltzmann entropy is an infallible indicator of phenomenal anti-entropic behaviour, nor that all such behaviour must be accompanied by decrease of Boltzmann entropy. In order to supply such a proof, we would need a clear method of cutting up the phase space of any model of classical mechanics into macrostates, and a demonstration that moving from a macrostate with a large volume to one with a small volume always constitutes an anti-entropic phenomenon. I do not think this is a hopeless project, but it is far from trivial and I will not attempt to carry it out here – nor would I know how to do it. Furthermore, in subsection 4.2.4, I will formulate a reservation towards this project. For these reasons, I propose not to adopt the decrease of Boltzmann entropy as a substitute for the phenomenal definition, because a lot of work has to be done to show its equivalence or non-equivalence. Also, and perhaps undeservedly, the Boltzmann entropy has played almost no role in the Maxwell's Demon literature. The only discussion of the demon I know of which uses lowering the Boltzmann entropy of a system as a criterion for demonhood is Albert 2000, [1]. The phenomenal criterion of section 2.1 has had a greater popularity, yet the literature has been truly dominated by references to the Gibbs entropy, to which we now turn. Once again, our question will be whether it can be molded into a substitute for the phenomenal definition of the demon.

2.3 Breaking the law: Gibbs entropy

2.3.1 Fine-grained Gibbs entropy

The Gibbsian conception of entropy is radically different from Boltzmann's conception, both mathematically and conceptually. For Boltzmann, a system is always in a well-defined state, and every state is associated with a value of the entropy. Gibbs, however, asks us to think not of a state, but of a *distribution*. Mathematically, it is a normalised density function on state space. Conceptually, there are several distinct ways of interpreting these distributions. One can

think of them as ensembles, infinite sets of equally prepared systems, where the density function gives the relative frequency of a state in the ensemble. Alternatively, one can consider them as mere mathematical tricks which allow us to calculate the behaviour of a single system; however, in this interpretation an equivalence between ensemble averages and properties of a single system would have to be proven, which is a complex and only partially solved problem. A third interpretation of the density functions is as measures of our subjective knowledge; they represent our uncertainty over the system's actual state. These interpretational questions will be ignored for the moment and the density function will be viewed as a measure of the probability for the system to be in a certain state without further specifying what is meant with 'probability'. A further discussion of this important subject is undertaken in section 2.4.

Consider a system T with a state space Γ and a normalised, possibly time-dependent, distribution function $\rho(\vec{x}, t)$ on Γ, where $\vec{x} \in \Gamma$ is a microstate and $t \in \mathbb{R}$ is the time-parameter. Then the **fine-grained Gibbs entropy** is defined as:

$$S_{FGG}(t) = S_{FGG}[\rho(\vec{x}, t)] = -k_B \int_\Gamma \rho(\vec{x}, t) \ln\{\rho(\vec{x}, t)\} d\vec{x}. \tag{2.2}$$

This entropy increases as the probability density becomes more 'spread out'. If $\rho(\vec{x}, t)$ is concentrated in one point, the entropy is $-\infty$, whereas a completely uniform density distribution over Γ yields a maximum. The fine-grained Gibbs entropy has some obvious advantages over the Boltzmann entropy, especially the fact that it does not depend on a partition of the state space into macrostates. Mathematically, it is much easier to apply, which is certainly one of the reasons it plays such an important role in statistical physics. On the other hand, it is conceptually less clear due to the use of density functions. But what really interests us is the question whether the fine-grained Gibbs entropy is an adequate and reliable indicator of anti-entropic phenomena. Does lowering the fine-grained Gibbs entropy always constitute a violation of the phenomenological criteria?

2.3.2 Fine-grained Gibbs and phenomena

It is instructive to look at the time-evolution of the fine-grained Gibbs entropy of an arbitrary distribution. Let T_t be the operator on Γ which represents the dynamical evolution of a system over a time t. We assume that this operation is bijective. Then the time-evolution of $\rho(\vec{x}, t)$ is given by:

$$\rho(\vec{x}, t) = \rho(T_{-t}\vec{x}, 0). \tag{2.3}$$

Accordingly, the time-evolution of the fine-grained Gibbs entropy is given by the following equation:

$$\begin{aligned} S_{FGG}(t) &= -k_B \int_\Gamma \rho(\vec{x}, t) \ln\{\rho(\vec{x}, t)\} d\vec{x} \\ &= -k_B \int_\Gamma \rho(T_{-t}\vec{x}, 0) \ln\{\rho(T_{-t}\vec{x}, 0)\} d\vec{x}. \end{aligned} \tag{2.4}$$

We apply a coordinate transformation $T_{-t}(\vec{x}) = y$, which is allowed since the time-evolution operator is bijective. The formula for the entropy now becomes:

$$S_{FGG}(t) = -k_B \int_\Gamma \rho(\vec{y}, 0) \ln\{\rho(\vec{y}, 0)\} [\frac{\partial T_{-t}(\vec{x})}{\partial \vec{y}}] d\vec{y}, \tag{2.5}$$

where only the determinant $[\frac{\partial T_{-t}(\vec{x})}{\partial \vec{y}}]$ is potentially time-dependent. But for a Hamiltonian system – and classical mechanics, to which we limit ourselves, is Hamiltonian – Liouville's equation tells us that a volume element is preserved under time-evolution; in particular, $d\vec{y} = dT_{-t}(\vec{x}) = d\vec{x}$, and therefore $[\frac{\partial T_{-t}(\vec{x})}{\partial \vec{y}}] = 1$. In other words, for Hamiltonian systems the fine-grained Gibbs entropy is constant in time!

Hamiltonian equations of motion conserve volume in state space, with an initial distribution (such as $\rho(\vec{x}, t)$) behaving like a incompressible fluid. The distribution can change its shape, but cannot contract or expand: here lies the reason for the fine-grained entropy's constancy. Thus, to lower the fine-grained Gibbs entropy of a classical mechanical system, the demon must itself be non-Hamiltonian: fine-grained demons must not conserve volume in phase space. But we want our demons to be models of classical mechanics; hence Hamiltonian; hence, conserving volume in phase space.

We want to have an entropy function which decreases when physical processes occur which violate the phenomenal second law. But the fine-grained Gibbs entropy is a constant of motion. It does not change with time. This is very strange, as it means that the evolution from a pressure difference to equal pressures, or a temperature difference to equal temperatures is not accompanied by a rise of fine-grained Gibbs entropy. Neither is there a logical implication from the creation of a pressure or temperature difference to a decrease of the fine-grained Gibbs entropy. In the light of these facts it may be somewhat surprising that the fine-grained Gibbs entropy has been perhaps the dominant definition of entropy used in recent literature on Maxwell's Demon – *prima facie* it is utterly useless. We will see later, in chapter 7, that it has not been applied in the straightforward way of using it as an indication of certain phenomena. Instead its main application has been in the Landauer-Bennett tradition where the fine-grained Gibbs entropy has been seen as 'distributed' over two distinct quantities: heat and information. In this context, the constancy of the entropy function is the very feature which makes it useful. But it is evident that since the fine-grained Gibbs entropy itself is not connected to the phenomena we are interested in, that essential connection has to be made by the distinction between information and heat. Whether this task can indeed be accomplished will be one of the central questions in the second part of this thesis.

For our present purposes, however, any further discussion of the fine-grained Gibbs entropy seems beside the point. Its value does not change, therefore it cannot act as an indicator of the occurrence of anti-entropic phenomena.

2.3.3 Coarse-grained Gibbs entropy

Because the fine-grained Gibbs entropy remains constant through time, another function has been thought up within the framework of the Gibbsian view. This is the coarse-grained Gibbs entropy, which unites certain characteristics of its fine-grained brother with the central feature of the Boltzmann approach: dividing state space into a countable (often, but not necessarily, finite) number of 'grains', being observable states. The idea is this. Suppose that the distribution starts in a small part of state space, concentrated entirely in one or perhaps a few grains. As time moves on, if the dynamics of the system are of an appropriate kind ('mixing'), it will become more and more spread out as different points in the initial distribution evolve in very different directions. Now suppose we

are restricted to observing the average distribution over a grain. What we will see from this coarse-grained perspective is a distribution that started out concentrated in a few grains evolving to a very evenly spread out distribution with more or less the same value in all grains. From a fine-grained perspective no such thing happens: on a finer scale (in the limit perhaps an infinitely fine scale) the distribution is still not uniform at all. But with our knowledge appropriately restricted, the distribution seems to reach near-uniformity.

Consider once again a system T with a state space Γ and a normalised, possibly time-dependent, distribution function $\rho(\vec{x}, t)$ on Γ, where $\vec{x} \in \Gamma$ is a microstate and $t \in \mathbb{R}$ is the time-parameter. Furthermore, there is a partition of Γ into denumerably many cells ω_i. The coarse-grained distribution function $\rho(i, t)$ is given by:

$$\rho(i,t) = \frac{\int_{\omega_i} \rho(\vec{x},t)d\vec{x}}{\int_{\omega_i} d\vec{x}} \tag{2.6}$$

$$= \frac{1}{V_i}\int_{\omega_i} \rho(\vec{x},t)d\vec{x}. \tag{2.7}$$

And the **coarse-grained Gibbs entropy** is defined as:

$$S_{CGG}(t) = S_{CGG}[\rho(i,t)] = -k_B \sum_i \rho(i,t)V_i \ln\{\rho(i,t)\}. \tag{2.8}$$

It can be shown that the coarse-grained entropy is always higher than or equal to the fine-grained entropy of the same distribution, where equality holds only if the distribution is uniform over all grains in which it has a non-zero value.

2.3.4 Coarse-grained Gibbs and phenomena

What can we say about decreases and increases of the coarse-grained Gibbs entropy? First of all, there is a clear limit on any lowering of a distribution's coarse-grained entropy, as it can never drop below its fine-grained entropy (which is constant for Hamiltonian systems, remember). So unless a coarse-grained demon is also a fine-grained demon, it cannot lower the coarse-grained entropy of a density function which is uniformly distributed over a certain number of grains. This means that a proper coarse-grained demon (where 'proper' indicates that it cannot lower fine-grained Gibbs entropy) can only work on non-uniform distributions in state space, which it has to rearrange into a uniform distribution over a smaller part of state space. It must, in other words, lower the amount of macroscopically distinguishable states over which the distribution has a non-zero value. Does this have a connection with anti-entropic behaviour? Answering that question can only be done after an interpretation of distributions has been chosen, which we will do in section 2.4. So we postpone a judgement of coarse-grained Gibbs entropy to subsection 2.4.2. Until and unless it turns out that the coarse-grained Gibbs entropy is a viable candidate for appearance in a definition of Maxwell's Demon – which it will not turn out to be – none of the proposed measures of entropy, neither the Boltzmannian nor the two Gibbsian definitions of entropy, is usable "as is" in this definition. Then, we will have to stick with the phenomenal characterisation of Maxwell's finite being.

2.4 Probability, distributions, ensembles

2.4.1 Ensembles and the demon

Our definition of Maxwell's Demon – given in subsection 2.1.2 – contains the clause that Maxwell's Demon has to produce anti-entropic effects 'with high probability'. Why would such a clause be necessary? Indeed, how could the notion of probability creep up in a discussion of systems from classical mechanics; after all, classical mechanics is deterministic and makes no use of probabilities. The current subsection will try to explain first how the introduction of ensembles of systems generates a need to speak about probability, and then why this introduction is justified and even necessary.

Maxwell's Demon will always be confronted with a single physical system in a single physical state. Unless we wish to give up the basic ontology of classical mechanics, there is little else we can believe it to come in contact with. Every system has one and only one state (even though we may not know it), and when the demon starts working on a system there is one and only one time-evolution which can take place: the time-evolution which follows from the initial conditions through the laws of classical mechanics. Therefore, the demon will either succeed or fail, and it is completely determined which of the two will happen. But suppose that we consider not a system with a single initial condition, but rather an ensemble of similar systems, all with different initial conditions. This ensemble could be described by a density function $\rho(\vec{x})$ on the basic system's state space, where $\rho(\vec{x})d\vec{x}$ is the fraction of systems in the ensemble with an initial function in the volume element $d\vec{x}$. Now for every system in the ensemble, the demon either fails or succeeds, and failure is completely determined by the initial conditions of the system. But looking at the ensemble as a whole, the demon may fail when working on some systems, and succeed when working on others. It is possible, for instance, that the demon succeeds in its entropy lowering labour for 80% of the systems, while failing for the remaining 20%. If the density function $\rho(\vec{x})$ is a reliable measure of the probability that the demon will come into contact with a system of a certain sort, we can say that the demon has an 80% chance of success, and a 20% chance of failure. Thus it becomes meaningful to demand of the demon that it succeeds 'with high probability' – in other words, 'for a large part of the ensemble'.

But why would we want the demon to be able to operate on an entire ensemble, and not just on a single system? Let's start by pointing out that it is trivial to think up a demon which can create a temperature difference for one given initial condition of a system. Just program the shutter normally operated by Maxwell's temperature demon to open and close on certain predefined times; for almost every initial condition of the gas[6] there is a sequence of openings and closings which will create the maximum pressure difference. We humans may not be able to find this sequence, but it nevertheless exists. Such a pre-programmed machine is certainly not forbidden by our laws of nature. Therefore, the question of accepting or not accepting the condition that a demon ought to operate on an entire ensemble is a very important one; if we decide against acceptance, demons can exist. We should demand that that neat-fingered being be able to create a temperature difference not just when confronted with one particular

[6]The exception are initial conditions which lead to an evolution in which some of the molecules never hit the shutter.

initial condition of the system, but for a large class of those conditions. But once again: why is this demand justified?

First of all, we might continue our argument, the number of systems which exist in the world, the state of which is very near to any certain specific initial condition, is probably very small if not zero. But what use has a demon if it cannot operate successfully for lack of systems with the right initial conditions? Such a demon could not be put to work; hence, it would not result in the creation of anti-entropic phenomena. Secondly, there is the problem of outside interventions. No system is completely isolated, so even if we have found or created a system with exactly the right initial condition – a remarkable feat –, interaction with the world outside would soon change this state. But this would spell disaster for a demon which works only on a very small set of initial conditions; even a small deviation destroys the possibility for the all too specific demon to operate. What use a demon if even the thermal movements of the atoms of the container walls, a kind of influence from the outside world which can hardly be prevented from changing the paths of the gas's molecules, will almost certainly stop it from operating? So what we need is a demon which can operate on a system *not* only when the systems has one single initial condition or a very small set of them; it should be able to operate successfully for a large class of possible initial condition, and, if at all possible, its operation should be robust under small external interventions.

This is where ensembles come in. We should interpret the initial ensemble as characterising the class of initial conditions with which the demon can be confronted. Imagine the following: a collection of infinitely many systems, where the distribution function $\rho(\vec{x})$ describes the frequency with which certain states appear. Now an infinity of demons, all faithful copies of the original demon, approach the ensemble, one to each system. They spit in their tiny claws, flex their neat fingers and start opening and closing shutters as needed. Some will succeed, others will fail: and it will depend on the ensemble what their average rate of success is and how well they succeed on average. A too-specific demon will fail almost all of the time, whereas a *real* Maxwell's Demon will succeed in almost every situation. To understand and hopefully quantify this, ensembles are needed; therefore, using ensembles is not only justified, but very important. This is reflected in the definition of Maxwell's Demon by the clause that it has to succeed with high probability.

2.4.2 Distributions and coarse-grained entropy

Now we return to the question whether we can replace the phenomenal definition with one which uses a quantitative measure of entropy. We already saw that the fine-grained Gibbs entropy was not a serious candidate, and the Boltzmann entropy was too besieged by problems to fulfill the role of definition without much more work being needed. This still left the coarse-grained Gibbs entropy as a possibility; to judge it, we need to find out what the distribution $\rho(\vec{x})$ actually signifies. So: what are these distributions?

One influential line of thought is the idea that the distribution is a measure of our subjective probabilities that a system is in one state or another. This does not seem to be a very useful idea in a discussion of demons, as in this interpretation, the demon would have to change our belief rather than the physical system – counterintuitive to say the least. Whatever the merits of this

idea may be, it is so at odds with both our phenomenal definition of entropy and the general approach taken in this thesis that it is safe to ignore it; for us phenomenal demonologists it can be of no use.

Another interpretation is that ensembles are measures of the frequency (in the infinite-time limit) with which a single system visits certain places in state space – in other words, the frequency with which the system attains certain states. But we are not interested in a demon which can change frequencies in the limit of infinite time; we want our demon to change heat into work right now.

The third major approach is to interpret the distribution as the characterisation of an *ensemble* of systems, an infinite, imaginary collection of systems where the distribution gives us the relative frequency in the ensemble of systems in a certain state. We have already seen in subsection 2.4.1 that such a 'space ensemble' – as opposed to a 'time ensemble', which is a collection of systems in subsequent states of a single system's evolution – is a necessary tool in defining what Maxwell's Demon actually is. If the Gibbs entropies are to tell *us* anything useful, we would make a wise choice in equating the distribution function which appears in them with the distribution function that describes an ensemble. Of course, I already rhetorically anticipated this equivalence when I named both of them $\rho(\vec{x})$.

Given this use of ensembles, we may try and find answers to the questions we still had concerning the Gibbsian entropy functions. The value of the fine-grained entropy is completely determined by (and always equal to) its initial value; and this initial value is something like a measure of the diversity within the ensemble. The more variety in initial conditions for the systems in the ensemble, the more spread out in state space $\rho(\vec{x}, t)$ and the higher the entropy. It was already clear that change in the fine-grained entropy could not tell us anything about the behaviour of a system, since it is nonexistent (whereas a system can, of course, exhibit some determinable behaviour). In addition, we can now see that the *value* of the fine-grained entropy doesn't tell us anything about a system either, but merely about the ensemble to which it belongs.[7] The coarse-grained entropy does not say anything about a single system either, as it too tells us something about the variety exhibited by the ensemble instead of the features of a single system. Where the fine-grained entropy is a measure of the overall variety in initial conditions, the coarse-grained entropy indicates the macroscopic variety, the distribution of systems over different macrostates (in the Boltzmannian sense). An army of identical coarse-grained demons would have to take, to lower the entropy, an ensemble of systems with widely different macrostates, and shape it into an ensemble where almost all systems are in a few macrostates. But this is hardly demanded of Maxwell's Demon. A creature which takes an ensemble concentrated in a single macrostate, that in which two containers of gas are at the same temperature, say, and transforms it into an ensemble of systems with all kinds of macrostates, in many of which there is a sizeable temperature difference between the two containers, would be a successful Maxwell's Demon. But it might not have lowered, but rather raised the coarse-grained Gibbs entropy. So a connection between the coarse-grained

[7] I would like to stress that this is a feature of the specific interpretation of the probability distribution which I am currently using. In other interpretations, the fine-grained Gibbs entropy *can* say something about a single system; but those interpretations do not justify the use of ensembles in the discussion of Maxwell's Demon.

Gibbs entropy and the phenomena we are interested in cannot be readily discerned. This is no proof that such a connection does not exist, but together with the fact that there has been little interest in it in the Maxwell's Demon literature it is my justification for ignoring the coarse-grained Gibbs entropy from now on. We will adopt, from now on, a phenomenal definition of Maxwell's Demon.

I will now try to elucidate what is hopefully the only remaining unclarity in our definition of Maxwell's Demon, which is the 'operating in a cycle' clause.

2.5 The cyclicity condition

2.5.1 Defining cyclicity

In subsection 2.1.2 I presented a phenomenal definition of Maxwell's Demon, which contained a clause about the demon having to operate 'in a cycle'. It is easy to show with an example why some such clause is needed. Suppose we have a system which has an internal source of work, such as a weight raised in a gravitational field, two volumes of gas at unequal temperatures or a chemical battery. Obviously, such a system could reliable produce any of the phenomena mentioned in the phenomenal definition of subsection 2.1.2: by using its internal source of work it can create a pressure or a temperature difference, or 'create' work from heat. But it cannot go on doing this indefinitely. The weight will reach the ground (or the centre of the earth), the temperature difference will decrease to zero, or the battery will run out of power. The system is not a valid demon because it 'cheats': it uses a hidden source of available energy to make it appear *as if* it creates available energy from useless energy. If we demand the system under scrutiny to operate in a cycle, we exclude such cheating: obviously, the height of the weight, the temperature difference between the gasses or the chemical potential difference in the battery is not the same at the start of operation as it is at the end. Using an internal source of work which is not replenished during the process implies that the system operates acyclically. Therefore, cyclicity is a sufficient condition if we wish to exclude such systems; imposing cyclicity on the demon ensures that it does not cheat.

But the requirement of cyclicity is in need of some further clarification, as it can be enforced with different degrees of severity. One possibility would be to require that at the end of a cycle the system must be in a state which is exactly identical in every physical detail to the state from which it started. This, however, would be quite unreasonable. There is most certainly no system in the world, and there will never be one, which operates in a cycle in this strict fashion. Using the strictest possible cyclicity condition, Maxwell's Demon will not exist in reality, but trivially so. The strict condition is too strong; we only have to forbid *relevant* changes, and can allow the demon to undergo *non-relevant* changes.

But what are the relevant changes? These, a first guess might be, include changes in two groups of properties: mechanical variables (such as the positions and speeds of macroscopic objects and the volumes of containers) and thermodynamical variables (such as pressures, temperatures and chemical potentials, which are only well-defined for macroscopic collections of particles). But this characterisation of relevant changes is at once too strict and not strict

enough. It is too strict in the sense that some mechanical variables should not be deemed relevant; for instance, changing the position of a block of concrete without altering its temperature, kinetic or potential energy, or any other of its original properties which might be used to do work, hardly seems 'cheating'. Yet it may not be strict enough because some microscopic alterations – which are neither mechanical nor thermodynamical changes – may stop the demon from functioning after one or a few cycles because they somehow interfere with its operation. Wouldn't this show that the demon is using a hidden source of energy, is somehow 'cheating'?

Therefore, two sub-conditions which we group under the name of cyclicity seem to be the following: the candidate-demon may not use reserves of readily usable energy stored within it (such as raised weights or batteries); and it may not undergo any changes that stop it from functioning in the future. When evaluating a candidate-demon we should not enforce cyclicity in the strict sense, but wonder whether the deviations from strict cyclicity – which are bound to take place – are irrelevant, or relevant. And a deviation is relevant if it uses a reserve of 'readily available' energy or endangers the demon's continued operation. These two kinds of deviation can be subsumed under a single heading. Confining ourselves to finite beings, cheating implies a breakdown in the future – at some point the weight will have reached its lowest position, the battery has run out of charge, etcetera. So the relevant condition to impose on the demon is that its operation should not cause changes which endanger its continued operation. Plugging these results into our phenomenal definition, we obtain:

Maxwell's Demon, clarified phenomenal definition: A sufficient condition for a system to be a successful Maxwell's Demon is that it can produce, with high probability and without endangering its own continued operation, at least one of the following phenomena without making any other changes in the environment:

- Two systems of the same temperature evolve to one system at a significantly higher and one at a significantly lower temperature.
- A significant amount of heat is converted completely into work.
- Two vessels containing the same gas at the same temperature and pressure evolve to one vessel with a significantly higher and one vessel with a significantly lower pressure, while the temperatures remains equal.

2.5.2 Is continued operation too restrictive?

In his book *Time and Chance* ([1], p. 109-110, footnote 12) David Albert questions the need for a cyclicity condition. He points out that the second law is broken when the total (Boltzmann) entropy of the system is lowered, and that whether its subsystems do or do not return to the macrostates they started in is irrelevant. In his usual flamboyant style, he writes:

> The first thing to say is that macroconditions of isolated systems which are overwhelmingly likely to lead to evolutions in the course of which the entropy of the system in question decreases are absolutely and unambiguously and straightforwardly in violation of the

> second law of thermodynamics – completely irrespective of whether any particular *component* of that systems happens to return to the macrocondition it *starts out* in or not – period, end of story.

But the story has not yet come to a close – not even for Albert, but certainly not for us. I suggested in the last subsection that the reason to enforce cyclicity was to exclude 'cheating' demons. Albert is in the happy position that he needs no such clause: false demons which use a battery or a raised weight in order to create temperature difference – think of a refrigerator – may be creating anti-entropic phenomena, but they are not lowering the Boltzmann entropy. (If, as Albert assumes, Boltzmann entropy turns out to be a good measure of entropic behaviour.) Cheating demons are not possible if the criterion used to judge them is the decrease of a reliable and quantitative measure of entropy. But we, who are not convinced that the Boltzmann entropy is as yet usable as such a measure, we, who rely on a phenomenal criterion, do need something along the lines of a cyclicity condition to exclude cheaters. The mere occurrence of an anti-entropic phenomenon can never prove that the demon is successful; if another phenomenon which 'raises the entropy' occurs at the same time, the demon might actually fail.

An alternative for the cyclicity condition is to enumerate all anti-entropic phenomena and define a demon as a system which can, with high probability, ensure that *at least one* phenomenon from the list takes place while at the same time *none* of the listed phenomena takes places *in reverse*. Creating a temperature difference is good, but not by exploiting an already existent pressure difference. Unfortunately, this alternative suffers from a fatal weakness: we do not have an exhaustive list of all anti-entropic phenomena, and it is not likely that we can construct one without begging important questions. For what criterion will we use to judge admittance on the list? It is certain that "Its reverse can be used to create phenomena on this list" can *not* be used, as it would indiscriminately disqualify *all* demons as cheaters. Perhaps a good criterion can be devised, but currently I am not able to imagine one.

What is true for cyclicity is also true for the clause that the demon must be able to operate continuously. As far as I can see there is no deep reason to be dissatisfied with a demon which would stop operating after a while, as long as we were sure it was not a 'cheating' demon. But we have no reliable criterion for cheating which is weaker than the requirement of continuous operation. This is the only reason I insist on using the requirement of continuous operation: not because it is intrinsically necessary for demons to keep going, but because it is the only way I have to exclude cheating demons. As a requirement it may be too strong, it may even be more vague than is desirable, but it is the best I have. It is a definite weakness in my account, and the price I have to pay for using a phenomenal definition. (For had I used a mathematical definition of entropy, I could just have demanded that the demon lower the total entropy of the world – not a problem in sight.)

This concludes our sketching of the portrait of the demon. None of the available measures of entropy was seen to be able to replace the phenomena in a definition of Maxwell's Demon. In addition, we highlighted and explained the importance of cyclicity – not to be adhered to in the strictest sense – and the use of ensembles, which introduces the notion of probability. Now that we have come to know it, we can turn to the necessity or contingency of the nonexistence

of this neat-fingered being. In chapter 3 we will look at an important argument for necessity; in chapter 4 I will present my case for the contingency of the demon's nonexistence.

Chapter 3

An argument for necessity

In this chapter, one simple but powerful and initially convincing argument for the necessity of the Second Law is presented. This argument tries to show that because the laws of classical mechanics are Hamiltonian, no successful demons – as defined in subsection 2.5.1 – can exist. I believe it to capture one of the main ideas of the Maxwell's Demon literature, an idea which forms the only half-articulated core of many more complex and specific arguments made by exorcists; it should be stressed, however, that I have nowhere found it in exactly the form presented here. I will therefore not attribute it to anyone. Chapter 7 will discuss arguments taken from the literature which I believe to be very close to the one presented here.

3.1 Heat and work in statistical physics

Let us first repeat the definition of Maxwell's Demon which we chose in the preceding chapter:

Maxwell's Demon, clarified phenomenal definition: A sufficient condition for a system to be a successful Maxwell's Demon is that it can produce, with high probability and without endangering its own continued operation, at least one of the following phenomena without making any other changes in the environment:

- Two systems of the same temperature evolve to one system at a significantly higher and one at a significantly lower temperature.
- A significant amount of heat is converted completely into work.
- Two vessels containing the same gas at the same temperature and pressure evolve to one vessel with a significantly higher and one vessel with a significantly lower pressure, while the temperatures remains equal.

In order to apply this definition to models of classical mechanics wherein a gas is presented by a multitude of tiny particles, we need to establish a connection between this atomistic description and at least one of the phenomena which occur in the definition. That is, we must make clear what 'temperature', 'heat' and 'work' mean in the context of classical mechanical models of statistical physical systems.

3.1.1 Temperature

The temperature of a gas, object or other collection of particles is related to the average kinetic energy of its constituents. For a gas in n dimensions ($n = 3$ in most realistic cases) which consists of identical particles of mass m, the correct formula is:

$$\overline{E}_{kin} = \frac{n}{2} k_B T, \tag{3.1}$$

with, of course, the kinetic energy of a particle being

$$E_{kin} = \frac{1}{2} m \vec{v}^2. \tag{3.2}$$

Very simple algebra yields:

$$T = \frac{1}{n k_B} m \overline{\vec{v}^2}. \tag{3.3}$$

Actually, this definition is not quite correct, as it implies that a system in which all the particles are at rest with respect to each other, but which moves very fast as a whole, has a high temperature. This is not the case: if all particles are at rest with respect to each other, the object's temperature is zero. Macroscopic movement does not count for temperature. The remedy for this problem is to stipulate that all $\vec{v}$'s are to be measured in the rest frame of the gas's centre of mass.

This is not a simple procedure which enables us to compute the temperature of any classical mechanical system. As a counterexample, imagine a system which consists of two equally heavy collections of particles which are both at zero temperature, but are moving away from each other. In the rest frame of this system's centre of mass, all particles have a kinetic energy; hence, the temperature would be positive according to our formula. In order to make a correct calculation, we would have to look at the subsystems individually, and observe them in their respective rest frames. But how does one carve up a system? If we do not carve it up into small enough pieces, compound systems such as the two balls are unfairly treated as one. But if we carve systems up in pieces which are too small, say individual molecules, our procedure will always yield a zero temperature. There is no easy answer to this problem; we will return to it later. For now, we'll confine ourselves to volumes of gas which we postulate can be seen as one single system for purposes of computing their temperature from the movement of their constituent particles. In effect, we assume that there is a clearly defined macro-realm (the gas, the container), and a clearly defined micro-realm (the individual molecules).

3.1.2 Heat and work

Energy occurs in two forms in thermodynamics, as heat and as work. As a first approximation to the truth, we can say that heat is energy in the form of molecular (or atomic) movement, whereas work is energy in a macroscopic form, such as the position of a weight in a gravitational field, or the velocity of a macroscopic body. Thus, the kinetic energy of a gas's molecules is heat, and the potential energy of a weight somehow attached to it is work.

Our definition of Maxwell's Demon says that a system is a demon if it can change heat completely into work. The 'completely' clause is meant to exclude

normal physical systems such as steam engines, which use the heat of hot water to create work. In order to do this, however, they must have access to a reservoir of cold water (or some other cold substance). By transporting heat from a hot object to a cold object, we can change part, but only part, of it into work. But it is not normally possible to change heat into work without making use of some temperature difference; without, that is, transporting part of the heat to a cold object. A being that can do this, that can change heat into work without sending part of it to a colder object, deserves to be called a Maxwell's Demon. It could make ships run on the heat of the ocean, no fuel needed.

So, something is a Maxwell's Demon if it can change microscopic kinetic energy into some macroscopic form of energy. Unfortunately, this definition is not as clear a one as we could hope for. Once again we are faced with a gap between microscopic and macroscopic that we have to draw ourselves – classical mechanics will not draw it for us. Which degrees of freedom are microscopic? Which are macroscopic? If ten billiard balls simultaneously and from the same direction hit a larger ball, transferring all their kinetic energy, have we transferred energy from a microscopic to macroscopic level? We are haunted by a problem identical to that which haunted us in our classical mechanical definition of temperature. We will adopt the same 'solution': "for now, we'll confine ourselves to volumes of gas which we postulate can be seen as one single system for purposes of computing their temperature from the movement of their constituent particles. In effect, we assume that there is a clearly defined macro-realm (the gas, the container), and a clearly defined micro-realm (the individual molecules)." Under these assumptions, heat and work can easily be defined.

3.2 Spinning the argument

Now we move to the argument for the second law's necessity which is to be the core of this chapter. We are working under the assumption – introduced in the previous section – that there are clearly separated microscopic degrees of freedom and macroscopic degrees of freedom. *Heat* is assumed to correspond to energy divided among the many microscopic degrees of freedom, *work* to energy contained in the macroscopic degrees of freedom.

The argument, which I will call the 'state space contraction argument', or *SSC* for short, proceeds in two steps. In the first, it is proved that completely changing heat into work corresponds to a contraction of the ensemble in state space. In the second, it is proved that a contraction of the ensemble in state space is impossible for a system of classical thermodynamics. If successful, the argument shows that at least one of the phenomena from our phenomenal definition cannot possible occur, given only the truth of classical mechanics. Since we claimed that temperature and pressure differences can easily be used to change heat into work, a proof that the latter effect cannot take place will force us to agree that the others cannot either. Hence, for as far as we've specified it, Maxwell's Demon is shown to be impossible if *SSC* succeeds. It would show that the second law is not contingent on classical mechanics, but necessitated by it. Let us turn to the argument.

3.2.1 Stage one: why contraction is necessary

Changing heat into work means transferring energy from microscopic to macroscopic degrees of freedom. Presumably, there are many more microscopic degrees of freedom than macroscopic ones. Suppose that we wish to lift a weight using the kinetic energy of a gas with $N+1$ particles. The number of macroscopic degrees of freedom is 1, that of microscopic ones $3(N+1)$.[1] This means that if we have a positive amount of energy E_0 to divide among the macroscopic degrees of freedom and zero to divide among the microscopic ones, there is only one possible state our system can be in. The weight is in the position in which it has an amount E_0 of potential energy, and all of the molecules lie still. But if we have a positive amount of energy E_0 to divide among the microscopic degrees of freedom and zero to divide among the macroscopic ones, the possible states of the system form an entire $(3N+2)$-dimensional plane in state space. Thus a much larger part of state space correspond with the second distribution of energy than with the first one.

Assume an $(N+1)$-particle gas with particles of mass m has a total kinetic energy (heat) E_h. What volume in state space do the possible states which exhibit this characteristic have?[2] For ease of calculation, we will assume the particles have a velocity in only one direction.[3] Let the particles be numbered 1 to $N+1$, and their momenta $p_1 = mv_1$ to $p_{N+1} = mv_{N+1}$. Then we have

$$E_h = \sum_{i=1}^{N+1} \frac{1}{2} m v_i^2 \tag{3.4}$$

$$= \sum_{i=1}^{N+1} \frac{1}{2m} p_i^2. \tag{3.5}$$

This is exactly the formula for the positive quadrant of the surface of an $N+1$-dimensional ball in position-momentum space with radius

$$r = \sqrt{2mE_h}. \tag{3.6}$$

The surface area of such a quadrant is

$$A = cE_h^{N/2}, \tag{3.7}$$

where c is a constant not relevant for our purposes. What *is* relevant is the insight that the volume in state space of all states with energy E_h scales as $E_h^{N/2)}$ – in other words, the volume in state space greatly increases as the kinetic energy distributed among the molecules increases. In sharp contradistinction, the number of states accessible to the weight if it is to have a potential energy of E_w is always one; there is one height and one height only at which the weight will have a particular potential energy. The total volume in state space accessible to our toy system – a container of ideal gas and a single weight we wish to

[1] There are an additional $3N$ degrees of freedom for the positions of all the N particles of the gas, but these do not bear any energy. We therefore neglect them.

[2] We use a velocity-position space with the standard measure on Cartesian space.

[3] This does not invalidate our conclusions. Using x dimensions ensures that in formula 3.7 the energy scales with $E_h^{x(N+1)}$ instead of $E_h^{(N+1)}$. My subsequent argument works for *any* $x \in \mathbb{N}$.

lift – is therefore proportional only to $E_h^{N/2}$, where E_h is the amount of energy distributed among the molecules; in other words, the amount of *heat* present in the system. It is thus easily seen that changing heat into work *decreases* the number of states the system can be in.

For the general case with $N+1$ molecules of gas and $M+1$ weights, the total volume in state space will scale with $E_h^{N/2} E_w^M$.[4] In addition, the total energy is fixed: $E = E_h + E_w$. With these two constraints, we can calculate at which value of E_h the volume in state space accessible to the system is greatest. The volume is $V = cE_h^{N/2} E_w^M$, with some constant c, which we can rewrite as

$$V = cE_h^{N/2}(E - E_h)^M. \tag{3.8}$$

Differentiating this equality to E_h, we obtain

$$\frac{dV}{dE_h} = cE_h^{([N/2]-1)}(E - E_h)^{(M-1)}[-ME_h + \frac{N}{2}(E - E_h)], \tag{3.9}$$

which leads to the following result for the maximum of V, which is also the *only* local maximum:

$$E_h = \frac{N}{N + 2M} E. \tag{3.10}$$

If $N + 1$, the number of molecules, is much larger than $M + 1$, the number of weights, – if, in other words, there are much more microscopic degrees of freedom than macroscopic degrees of freedom – then $\frac{N}{N+2M}$ is almost unity. The maximum volume in state space is reached for an energy distribution among the gas and the weights in which almost all energy is possessed by the gas; and the volume decreases monotonically as the energy distribution is removed farther and farther from this maximum. This proves that no significant transfer of energy from microscopic to macroscopic degrees of freedom can ever take place without (greatly) decreasing the amount of accessible states. At least, no such transfer can take place if $M \ll N$, a supposition which we will scrutinise in section 3.3.1.

3.2.2 Stage two: why contraction is impossible

We will now show why contraction of an ensemble in state space is impossible. Actually, this proof is very simple, and we will need to spend more time showing why it is relevant than proving it. We recall from 2.3.2 our exposition of the constancy in time of the fine-grained Gibbs entropy. The most important ingredient of this discussion was the fact that classical mechanics is Hamiltonian, and that Liouville's theorem shows that in a Hamiltonian system, volumes in state space cannot contract or expand during time evolution. Recalling this is giving the proof we need. Classical mechanics is Hamiltonian; hence, an ensemble in state space cannot possible contract. Hence, by the results of the previous section, the kinetic energy of a gas cannot ever be used to raise a weight. Such a process would correspond to an enormous decrease of the ensemble's volume in state

[4] It can be easily seen that the volume in state space accessible to a system of $M+1$ weights in a uniform gravitational field scales with E^M: the accessible states form an M-dimensional plane which is stretched by a factor a in every direction when the energy is increased with a factor a.

space, and this is absolutely forbidden by classical mechanics itself. Therefore, the validity of the second law is necessitated by classical mechanics. QED.

Well, perhaps not quite. First of all, we should notice that our argument does *not* prove that it is impossible that a system starting in a state with much molecular kinetic energy evolves towards a state with less molecular kinetic energy and more energy stored in the potential energy of the weight; in other words, we have not proven – and could not prove, since it is false – that heat cannot be changed into work. The fact that classical mechanics is Hamiltonian does not, as such, forbid the evolution from any particular initial state to any particular final state. Such an evolution is not accompanied by a change of volume in state space, since at any moment there is merely one state. Volumes, Hamiltonianism and *SSC* only come into the picture when we start looking at ensembles.

We can return to subsection 2.4.1 for arguments for the necessity of using ensembles. The demon is supposed to operate on a large class of systems, that is, on all the members of a big ensemble. If we present to it a container full of gas at a certain temperature, we want it to be able to do its job, and succeed with a high probability. We do not want the demon to fail when presented with almost any initial configuration; we want it to do its job reliably. Of the entire ensemble, it must take the great majority of systems to a final state where work has been done using only the gas's heat. This criterium for success means that the combined system of the demon and the gas must evolve from an initial ensemble where the gas is spread out over a large part of state space and the demon is in its (single) initial state, to a final ensemble where the gas is spread out over a significantly smaller part of state space. And because classical mechanics is Hamiltonian, this is *impossible – unless* the demon can end up in many different states, unless, that is, its ensemble expands in state space to make up for the decrease of the gas's. The demon can only work if it has the possibility to end up in many different states when presented with different initial states of the gas.

This is where the cyclicity condition of section 2.5 enters. We settled there for the demand that the demon do not endanger its own continued operation; that, in other words, its final state be, with high probability, a state from which it can still function. This demand now translates into the following requirement: the great majority of the states the demon can end up in, must themselves be possible initial states from which the demon can work when presented with a gas. Unfortunately, this is impossible. For it would mean that in the joint ensemble of all the possible initial demon states and all the possible initial gas states, the great majority of the states would end up, after evolution, in a much smaller ensemble, namely that of the possible final demon states – more or less identical in size to that of the possible initial demon sates – combined with that of final gas states – much fewer than possible initial gas states, if the demon is to be successful. Algebraically, where D_i is the volume in state space of the demon's initial ensemble and the other terms should be clear, this can be expressed as follows: $D_i = D_f$, $G_i \gg G_f$, and therefore $D_i G_i \gg D_f G_f$. But this is impossible, since Liouville's theorem tells us that $D_i G_i = D_f G_f$. This algebraic equation should not be taken too literally, since in general the final ensemble can not be written as a product of a Demon-ensemble and a Gas-ensemble – there will be correlations between the state of the gas and that of the demon. Having a product ensemble as final ensemble is, however, the

best possible outcome for the demon; any other form of the final ensemble only increases the diversity of its final states. By *reductio ad absurdum*, then, we have shown that if the demon can end up in many possible states, the great majority of these cannot be good initial states for it to continue operating. So, if the demon is to be successful, it endangers its own continued operation. But to be successful, the demon may not endanger its own continued operation. Contradiction. Conclusion: the demon is not successful.

Please note that we nowhere assumed that the demon is 'in an ensemble', or that it must 'confront an ensemble of systems'. The demon is a single system in a single state, working on a gas in a single state. But what we did show is that *if* the demon can operate both reliably and 'cyclicly', *then* the ensembles we constructed must behave in a way which they cannot possibly behave. The ensembles are not to be interpreted realistically, they are a way of presenting our requirements mathematically and deriving a contradiction from them.

One may worry about the asymmetry of the present argument. If it shows that Maxwell's Demon cannot lower the entropy, should it not also show that it cannot raise the entropy? The answer to this question is negative: taking the states of a small part of state space via Hamiltonian evolution to all the states of a larger part is, of course, impossible. But taking all the state of a small part of state space via Hamiltonian evolution to a subset of a larger part of state space is quite feasible – and it is enough. The State Space Contraction argument is asymmetric in exactly the same way as an argument showing that you cannot put a bucketful of water into a teacup. You *can* put a teacupful of water into a bucket; it just won't fill it up.

This completes *SSC*. As far as I am able to tell, the argument is sound. It makes no illegal use of the cyclicity condition, and it does not use illicit ensembles. If its presuppositions can stand the test of criticism, it shows that the second law is necessitated by classical mechanics.

I will, of course, argue that they cannot.

3.3 Counterarguments

In this section, two counterarguments to *SSC* – the State Space Contraction argument – will be presented. The first calls into question the assumption that there must always be more microscopic than macroscopic degrees of freedom. It is developed in subsection 3.3.1, and I claim it defeats *SSC*. Indeed, because no physical differences between microscopic and macroscopic degrees of freedom seem to follow from classical mechanics, it is suggested that it is impossible to derive the second law from classical mechanics. The second counterargument calls into question the assumption that the dichotomy between microscopic and macroscopic degrees of freedom is justified and illuminating in the present discussion. This counterargument is introduced in subsection 3.3.2, but discussed in the next chapter, where it will be elaborated in an attempt to paint a lucid picture of the second law and its relations with classical mechanics, Hamiltonianism and the constitution of matter.

3.3.1 Macroscopic multiplicity, microscopic sensitivity

We assumed in subsection 3.2.1 that the number of microscopic degrees of freedom is much greater than that of macroscopic degrees of freedom. This assumption was absolutely crucial, since only by setting $M \ll N$ in equation 3.10 could we conclude that any significant transformation of heat into work would be accompanied by a decrease of the ensemble's volume in state space. And only thus were we able to infer that Maxwell's Demon would need to break Liouville's theorem, which is of course impossible. This assumption has been called into question by David Albert ([1], p. 107 and further), whom we already met in subsection 2.5.2. Why couldn't we have a system with *many* macroscopic degrees of freedom? For instance, why can't we have a system which consists of a container with N particles of gas and much more than $3N$ heavy, rigid weights suspended in a gravitational field? Because there are less microscopic than macroscopic degrees of freedom, with a sufficiently ingenious mechanism we might be able to convert a very sizable portion of the gas's kinetic energy into potential energy. This mechanism should ensure that different initial states of the gas will result in different final states of the weights – that's all. Let us call it the Too Many Weights Machine, abbreviated *TMWM*. Against this contraption, *SSC* is powerless.

There can be several responses to this counterargument. A first, very bad, one is pointing out that as a matter of fact there are much more molecules in even a modest amount of gas than there are weights available to us. This is true, but indeed *as a matter of fact* – in other words, contingently. Classical mechanics does not tell us anything about the number of weights in the world. This response is irrelevant, as is any response like it.

A second line of attacking the current counterargument starts by noticing that the *TMWM* has a very special property, namely *microscopic sensitivity*. Albert defines it as follows:

> And we have learned something interesting (by the way) about what *sorts* of physical things those demons are going to need to *be*. They are going to need to be *microscopically sensitive*: they are going to need to be capable of tailoring their *macroscopic behaviors* to the particular *microcondition* of the system they happen to be *operating* on; they are going to need to be the sorts of systems (that is) whose final macrocondition is not predictable, and not even *approximately* predictable, from the initial macrocondition of the larger isolated system of which it forms a part.[5]

We can go on to ask whether there might not be some reason or another why microscopically sensitive devices cannot exist. Unfortunately for this line of thought, it is ridiculously easy to think up a model of classical mechanics which is microscopically sensitive. Imagine, for instance, a container, the vertical walls of which consist of many (much, much more than N) very thin rods of metal which extend to such a height that they have to be considered macroscopic objects. Since they are standing on their ends, they are in a highly unstable equilibrium position, and whenever a gas molecule hits one of the rods, it absorbs (part of) the molecule's kinetic energy and falls over. As it reaches the ground,

[5] Albert 2000, [1], p. 108-109.

it hits a ball to which it transfers its energy, and the ball merrily rolls away into infinity over a frictionless plane. Obviously, this device's final macroscopic state is highly dependent on the initial microcondition of the gas. Still, we may be on to something here, and there are some tricky questions left to ask. Maybe all microscopically sensitive devices must break the 'continued operation' clause? It seems, at first glance, that the container with the walls of rods breaks this demonic requirement – after all, its walls disintegrate as the rods fall down –, and perhaps all microscopically sensitive devices must. Is there a reason to suspect that the *TMWM* cannot continue to operate after changing heat into work one or several times successfully? The only difference between the *TMWM* after several successful runs and the *TMWM* before is that there is a larger amount of energy divided amongst its weights. Reconsidering formula 3.10:

$$E_h = \frac{N}{N+2M}E, \tag{3.11}$$

we see that if

$$E_{weights} > \frac{2M}{N}E_{gas}, \tag{3.12}$$

we are already past the maximum and we would actually *decrease* the ensemble's volume in state space by converting heat into work.[6] If the ratio of the weights' energy to that of the gas exceeds the ratio of twice the number of macroscopic to the number of microscopic degrees of freedom, the $TMWM$ stops working. There is, however, an easy remedy for this problem: increase the number of weights. By making it arbitrarily large, we can go on changing heat into work arbitrarily long; and if there are many more macroscopic weights than microscopic particles in our universe, the clause of continued operation is only broken in an uninteresting way (namely, when there is no more heat).

It is time to make an important observation. We have divided the world into microscopic and macroscopic degrees of freedom, but we have not introduced any physical differences between the two except for size. Furthermore, classical mechanics is scale invariant: 'size does not matter'. *Relative sizes* might matter, but in constructing our demons we have not yet seen that they do, and I know of no demonstration. The important observation is that *if* there are no physical differences between microscopic and macroscopic degrees of freedom, *then* we cannot possibly derive an asymmetric relation between them. If microscopic and macroscopic degrees of freedom are only different in physically uninteresting ways, there cannot be a second law. So if there is a second law, a necessary ingredient for deriving it is some law of nature or some fact about the world which establishes an asymmetry between microscopic and macroscopic degrees of freedom. Classical mechanics does not seem to supply us with such an asymmetry. Therefore, almost as a matter of logic, the second law cannot be derived from classical mechanics.

Perhaps all microscopically sensitive devices in our universe will be defeated because – as a merely contingent fact, given only classical mechanics – all macroscopic objects consist of many microscopic objects. There are always more microscopic degrees of freedom in our world than macroscopic ones, because our large objects are made of small atoms; and the interaction between the

[6] These formulae are for a 1 dimensional gas, but changing N to nN makes them valid for an n dimensional gas.

molecules of the gas and the atoms of the *Too Many Weights Machine* may defeat the contraption. This line of thought might be combined with the exorcisms of Chapters 5 and 6 to argue that the *TMWM* cannot exist in our world. But presently, we seem to have defeated the State Space Contraction argument. The fact that classical mechanics is Hamiltonian can never prove the second law, because it does not establish an asymmetry between microscopic and macroscopic degrees of freedom.

3.3.2 Dissolving the microscopic/macroscopic dichotomy

In the previous subsection, *SSC* was defeated from within the conceptual scheme it had established, a conceptual scheme where the world was divided into the microscopic and the macroscopic, each with its own distinct phenomena. The microscopic world stores its energy as heat, the macroscopic world as work. But we already saw in section 3.1 that this distinction is not unproblematic. Firstly, there are no clear cut boundaries between the microscopic and the macroscopic world (at least not in classical mechanics in general, though these boundaries may be easily drawn in our particular world). Secondly – something upon which we did not touch before – there is no *a priori* reason to assume that the world falls apart into two realms instead of three, four, five, seventy-two, or infinitely many. Neither does classical mechanics furnish us with such a reason. Recognising and exploiting especially this latter reason for rejecting the presuppositions of *SSC* as begging the question, will lead us to a simple argument which shows that not only is the second law contingent given only classical mechanics, but we do not even need microscopically sensitivity devices in order to change heat into work. This argument will be presented in the next chapter.

Chapter 4

Matters of scale and contingency

We have seen in Chapter 3 that the State Space Contraction argument fails to establish a necessary connection between the second law and classical mechanics. The reason that it failed to do so is that classical mechanics as such introduces no relevant asymmetry between microscopic and macroscopic degrees of freedom. We also deduced that every Maxwell's Demon ought to be microscopically sensitive, in the sense that its final macrostate must depend very sensitively on the gas's initial microstate; and that we need an ungodly amount of weights or other macroscopic objects in order to drain even a modest amount of gas of its kinetic energy, because the number of macroscopic degrees of freedom must exceed that of the microscopic ones. Together, these conclusions suggest that the reason that the second law seems to hold in such a strong sense in our universe has something to do with the fact that we cannot build macroscopic objects which do not consist of microscopic ones; perhaps with the fact that all interactions between macroscopic and microscopic bodies in our world are through the interaction between those microscopic bodies which are free and those which are the constituent parts of the macroscopic ones. Such suggestions will be looked at in Part II of this thesis.

In this chapter, we will take a closer look at matters of scale, and I will argue that microscopic sensitivity is not needed to defeat *SSC*. Maxwell's Demon *could* use zillions of weights, but does not *have* to do so. I will demonstrate once more that the second law is not necessitated by classical mechanics, but in addition I hope to indicate the kind of facts (or laws?) that the second law is contingent on in our universe.

First, in section 4.1, we will momentarily leave the perspective we've built so labouriously in the previous chapters, and look at four machines which will lead us to important insights concerning the nature of the second law. A model of classical mechanics is constructed which is a successful Maxwell's Demon but is not microscopically sensitive and has only very few 'macroscopic' degrees of freedom. As a prelude to this construction, three other machines will pass in review: Smoluchowski's famous one-way valve, which fails to break the second law because it is 'doomed by fluctuations'; a macroscopic equivalent of the one-way valve which *does* work – although it does not break the second law in a strict

sense; and a macroscopic equivalent of the one-way valve which does not work. Together, these four models set the stage for a discussion in section 4.2 of the role different scales play in thermal physics, and how the validity of the second law is tied up with considerations concerning these scales. The last section, 4.3 concludes Part I of the thesis by affirming the contingency of the second law, and summarising the kinds of fact it is contingent on.

4.1 Four models of the pressure demon

4.1.1 Smoluchowski's one-way valve

Now we embark on an exposition of several models – easy to visualise and 'get a feel for' – which will give us insight in the subtleties of the second law. For the moment, we will leave the world of volumes in state space behind, and focus on thermodynamical quantities such as temperature, pressure and heat as they appear in simple systems of classical mechanics. In section 4.2, we will project our conclusions back into the language of state spaces, Hamiltonianism, and so forth.

We start by discussing one of the most famous models from the literature on Maxwell's Demon, Smoluchowski's one-way valve, and his refutation of this design. Smoluchowski 1912 [30], one of the early classics of the exorcist literature, contains a very short description of perhaps the simplest example of a mechanical demon: a one-way valve. Take two containers of gas at equal temperatures with a little hole between them, and instead of a shutter, put a valve on it which opens only one way, say into B. The easiest way to visualise the valve is as a little trapdoor held closed by a very weak spring. Every time a molecule in A hits the trapdoor, it opens (we assume the spring is very weak) and lets the molecule pass. But whenever a molecule from B hits the valve, it remains shut, because it is pressed against the container wall. In this way, a pressure difference between A and B will build up without any change in the environment or the demon taking place, which qualifies the valve as a Maxwell's Demon.

Smoluchowski, however, does not accept this *prima facie* correct reasoning. He points out that the valve will be at the same temperature as the gas. That means it will be subject to thermal fluctuations: its kinetic energy will randomly change by amount of – on average – $\frac{1}{2}kT$. There are two possible scenario's, depending on the strength of the spring which holds the trapdoor closed. Either the spring is so strong that the thermal fluctuations can hardly get the valve to move; but in that case, no molecules will be able to pass either. (After all, the kinetic energy of the molecules in any direction is also of the order kT.) Or the spring is so weak that molecules from A can push aside the trap door and enter B, but then the door will be a constant victim of thermal fluctuations which make it jump one way, then the other, and stop it from functioning as a pressure demon. Thus, the fact that the trapdoor has the same temperature as the gas, and therefore the same average kinetic energy fluctuations, defeats Smoluchowski's design.

One may wonder whether the trapdoor would not still function as a demon, if somewhat less efficient than originally envisaged. It is open part of the time, in which case it does nothing; but it is closed another part of the time, in which

case it *does* work. Therefore, it would create at least somewhat of a pressure gradient, although not as efficiently as we could have hoped. Bennett 1987 [3] attempts to remove this objection. Indeed the trapdoor will stop some molecules in B from entering A (namely when it is closed). But this is compensated for by the fact that every time it bangs shut, any molecule in B which is in its path will actively be pushed into A, and molecules just passing from A to B may be bounced back. An analytical solution of this problem is quite involved, but computer simulations with a similar valve have confirmed Bennet's result; see Skordos & Zurek 1992 [29]. It appears that Smoluchowski's victory over this particular demon is final – I know of no authors who wish to contest it.

4.1.2 Macroscopic gas: a demon that works

I suggested in subsection 1.3.2 that the demon was used by Maxwell to show that the validity of the second law hinges on the fact that neither we nor any tools we can make have the ability to 'trace' and 'seize' molecules at will. In order to better understand this inability of ours, it is helpful to examine a situation in which we *can* trace and seize 'molecules' at will – better yet, a situation in which a mechanical quasi-demon does this for us. So let us examine a *macroscopic* model, modeled after Smoluchowski's microscopic one. We will make something akin to, but not quite the same as, a scale-transformed version of the one-way valve.

Imagine two very large containers A and B of equal size, placed in space far enough from heavy bodies to make gravitational forces negligible. A multitude of heavy, hard, macroscopic balls continually bounce through these containers, colliding with the walls and each other in a dissipationless and nonelastic way. Initially, the balls are divided evenly among the two containers. But a one-way valve is present in the wall connecting them. This is a trapdoor consisting of a rigid metal sheet normally closed by a spring weak enough that the energy needed to open the trapdoor is much smaller then the average kinetic energy of the balls. If a ball from A hits the door, it can push it open against the force of the spring and enter B. But a ball from B will only succeed in pushing the rigid metal sheet against the equally rigid container wall – it cannot enter A.

Thus, more balls will go from B to A then the other way around. The number of balls in A will fall, and that in B will rise. If we install a tube between A and B with a turbine in the middle, this difference can be exploited by having the turbine raise a weight whenever it is turned in the preferred direction. The net effect of the machine will be that the average kinetic energy of the balls is lowered and a weight is raised. In a sense, but *only* 'in a sense', we have a Maxwell's demon at work.

Three complaints may be raised against the claim that we have just constructed a Maxwell's Demon at work. The first is, of course, that we have not really changed heat into work, but have changed the kinetic energy of macroscopic balls into potential energy. This is true, and it is the reason I said we only have a Maxwell's Demon in a certain sense. But whether the distinction between heat and kinetic energy of macroscopic balls is really so fundamental, is a question which will receive more attention in section 4.2. The second complaint is that we have neglected the fluctuation phenomena which must defeat this machine as they defeated Smoluchowski's one-way valve. Actually, this complaint is unjustified. The machine described above works, and it is interesting

to point out the differences between it and Smoluchowski's demon which make the analogy break down. We will do this shortly. The third complaint is that the net effect of the macroscopic demon is more than what has just been stated: there must also be a heating up of the trapdoor, the spring or the container walls. This complaint is justified. I'll come back to it after I've discussed the previous one.

The only interactions between the balls and the metal door are the collisions which happen when either a ball in A or one in B hits the door. If a ball in B hits it, nothing happens to the metal sheet: it stays in places, is completely rigid and there is no energy exchange between the ball and the sheet. If a ball in A hits the door, something does happen: it opens. A small amount of energy, ϵ, is imparted first as kinetic energy to the door, then as potential energy to the spring, then once again as kinetic energy to the door until it bangs shut, at which point the energy is dissipated as heat in the door and the container wall. When no collision is taking place, the metal sheet has no kinetic energy, except for that given to it by random fluctuations in accord with its temperature. But for every temperature below the metal's melting point, we can be quite certain that the average kinetic energy of the sheet will be much less than the average energy of the macroscopic balls! Macroscopic objects in our world do not experience thermal energy fluctuations on macroscopic scales – one's furniture never spontaneously moves around. Hence, the macroscopic machine we are talking about is not defeated by fluctuations.

To return to the third complaint, it is indeed true that a process of dissipation is taking place whenever the door bangs close against the container wall. This defeats our demon if and only if this breaks the cyclicity condition in some way. We'll return to that issue in subsection 4.2.3. For now, notice one important fact: the demon *does* succeed in the sense that it uses the kinetic energy of the balls to do work. To find out what features of the machine enable it to do so, what makes it different from Smoluchowski's device (except for its size) we'll consider another macroscopic model: one which does not work.

4.1.3 Macroscopic gas: a demon that fails

Imagine the very same machine, but with this difference: instead of rigid steel plates the containers consist of metal balls held together with springs. These balls are exactly the same size as those which are to be used to do work. Whenever one of those hits the wall, it exchanges energy with the balls in the wall. In the equilibrium situation all balls in the entire machine have the same average kinetic energy; there is no energetic difference between those that are held fast by springs and those that fly through space. In addition, the machine is fitted with a one-way valve, which is still made of a normal rigid metal sheet and a spring, as well as a turbine, a weight, and so forth. This contraption does not work.

It does not work because the one-way valve can no longer operate. In order to let molecules pass from A to B, the energy needed to open the trapdoor must be far less than the average kinetic energy, $k\mathcal{T}$, of the balls: $\epsilon \ll k\mathcal{T}$. But then the trapdoor cannot be at rest against the container wall, since this wall consists of balls with an average kinetic energy of $k\mathcal{T}$. When in contact with the wall, the sheet will experience energy fluctuations of order $k\mathcal{T}$, and thus be defeated. But if the sheet cannot come into contact with the wall, it cannot

dissipate its own kinetic energy after a collision – and it will be defeated too. To make things worse, imagine that the trapdoor itself and the spring which constrains it consist of the kind of ball that the rest of the machine is made of. If they come into contact with their surroundings – and in order to operate they'll have to – all these balls will soon have an average kinetic energy of kT. So if the spring is weak enough to allow any balls to pass at all, it is also too weak to keep the door in place. It is clear that this trapdoor is defeated by the same phenomena that defeated Smoluchowski's one-way valve.

Indeed, it cannot but be defeated by the same phenomena, since it is a faithful copy of the one-way valve. It is Smoluchowski's machine, after a scale transformation. (All right, we have changed inter-atomic forces to springs; but the idea is the same.) In Smoluchowski's device the little trapdoor consists of atoms, the walls of the containers consist of atoms, and the particles whose kinetic energy are to be exploited to do work are atoms too.[1] The situation is analogous in all important respects to that of our second macroscopic machine, where the trapdoor consists of balls, the walls of the containers consist of balls and balls are the particles to be manipulated too – only the scale is different. Because any interaction with a ball that is flying around involves energies of order $k\mathcal{T}$, the trapdoor must be sensitive to those energies – and preferably to ones quite a lot lower too. But at the same time, *every* interaction involves energies of order $k\mathcal{T}$, so if the trapdoor is in contact with its environment – which it must be – it constantly experiences fluctuations large enough to prevent its operation.

As already remarked, the present contraption is a scaled up version of the original one-way valve. We know that classical mechanics is scale-invariant, and therefore, what is true about Smoluchowski's device is just as true about the present macroscopic one. But that being said, it is immediately clear that we can create a working Maxwell's Demon which changes atomic motion into potential energy by *scaling down* the machine from subsection 4.1.2. We will do that now.

4.1.4 The tinyon machine

Suppose that a new class of particles, *tinyons*, is discovered which are much smaller than atoms; the difference in size between a tinyon and an atom is the same as that between a molecule and one of our macroscopic balls. These tinyons behave almost the same as molecules: they can be used to create rigid sheets of tiny-metal and all other kinds of rigid structures. In short, they can be used to make molecule-size rigid tools in exactly the same way that atoms and molecules can – and macroscopic balls held together by springs cannot – be used to make macroscopic rigid tools.

If the containers, the trapdoor, the spring of the trapdoor, the turbine and the weight are all created of tinyons, and the containers are filled with molecules, we have a situation analogous in all relevant respects to that of our first macroscopic machine. Since classical mechanics is scale invariant, there can be little doubt that it works, as everything we said about that big machine must also be true about this new, small machine, given the appropriate substitutions. But that means that the kinetic energy of molecules can be exploited to do work. I

[1] Well, molecules. The difference is small enough not to matter.

repeat: if we can build tools of tinyons, we can change heat into work; just as we can change the kinetic energy of macroscopic balls into work because we can make tools of atoms and molecules. The reason such a thing does not happen in reality is not that everything must fluctuate; it is not that measurements have an intrinsic entropy cost; it is not that memory erasure has an intrinsic entropy cost – as the exorcists we'll meet in Part II would have us believe. The reason that heat cannot be transformed into work is merely that there are no such things as tinyons.

But we're going too fast. There was dissipation in our first macroscopical model, and there will be dissipation in the tinyon machine too. Does this not defeat the demon after all? It is time to return to our previous terminology, our definition of Maxwell's Demon, considerations of volume in state space, and use the imagery of the section we now bring to a close to round off the arguments for the second law's contingency.

4.2 Scale and thermal physics

4.2.1 Temperature or temperatures?

We retrace our steps to section 3.1, where we took some pains to define temperature within the context of classical mechanics. We had to assume that the world could be carved up into definite macroscopic objects, and temperature was then defined using the average kinetic energy of the constituent particles of such a macroscopic object relative to the rest frame of the object's centre of mass. But, as we have seen, classical mechanics is scale invariant, and there is no reason to restrict this definition to the 'microscopic' and 'macroscopic' levels as we generally define them. Nor *is* the definition of temperature so restricted. For instance, astronomers regularly speak about the temperature of star clusters. This temperature is a measure of the average kinetic energy of the stars in the star cluster, in the reference frame defined by the cluster's centre of mass. The temperature so defined has nothing to do – nothing at all – with the temperature of the stars; the latter being a measure of the average kinetic energy of the atomic and subatomic particles that are the constituents of the star, measured in the rest frame of the star's centre of mass. In the same way, two different temperatures can be ascribed to our macroscopic machine from subsection 4.1.2: one is a measure of the average kinetic energy of the big balls, the other of the average kinetic energy of the atoms which make up both balls, container and trapdoor-mechanism. These temperatures can be – indeed, will be – wildly different; and this difference allows the machine to operate.

There is no such thing as temperature. Or, to say it more clearly, there are many temperatures. At each *scale* we can introduce a new temperature; and if there were no recognisable scales in the world but a continuity of sizes of natural objects ranging from the infinitely small to the gigantically huge, there would be no natural measures called 'temperature' at all. And, by the same token, no such things as 'heat' and 'work'. The validity of the second law thus hinges on the fact that in our world it is – so it seems – possible to identify 'natural scales'. Why is that possible? To see how intricate the answer to this question can be, we only need to return to our example of the star cluster: the fact that there are such easily definable things as 'star clusters' which are made

up of objects that all have roughly (very roughly) the same size, is presumably a result of the law of gravitation, certain facts about the Big Bang, and many complicated and perhaps poorly understood details of star formation and cluster formation. It is certainly not a consequence of anything as clean and simple as the Hamiltonianism of our world!

So if the second law is to hold, there must be natural scales to make the terms 'heat' and 'work' *meaningful*. But this is not enough the establish its validity, as we will see in the next subsection.

4.2.2 Scale non-invariance

Every argument for the second law which makes no use of special facts concerning scales in our universe, is faced with the following inescapable dilemma: it does not establish the second law as saying that the kinetic energy of atoms and molecules can never be changed into work; it rather establishes a scale invariant version of the second law which claims that kinetic energy at *any* scale cannot be transformed into work. But the scale invariant second law is simply false. Its falsity is illustrated by the macroscopic machine of subsection 4.1.2. This contraption transformed the kinetic energy of macroscopic balls into work. Thus, the second law did not hold on that scale. If classical mechanics were the only relevant part of physics, this would simply imply that the second law does not hold on any scale.

So how is it possible that the second law *does* hold on some scales – namely, the scale where temperature is used as a measure of the kinetic energy of molecules and atoms? The four models we reviewed furnish the answer to this question: we saw that if the container and the trapdoor were made of particles on a scale *lower* than that of the particles we wished to take the heat from, the machine worked; this was the case with both the first macroscopic machine and the machine using tinyons. And we saw that if every part of the machine was made of particles on the same scale as those we wished to take heat from, the machine did not work: this was made plain by both Smoluchowski's original device and my macroscopic equivalent. The explanation I gave was that the trapdoor must fluctuate with smaller energies than the kinetic energies of the particles that it has to stop or let through; and this is only possible if its constituent particles are either at a comparable but lower temperature – in which case they will warm up, which will defeat the demon – or are at an incomparable temperature, at a lower scale.

But we already have the tools to understand this fact in a manner abstracted from the specific operations of the one-way valve. We know that classical mechanics is Hamiltonian, and that this means that volume in state space is preserved. In chapter 3 it was shown – assuming that there were only two scales, dubbed 'microscopic' and 'macroscopic' – that under the constraint of Hamiltonianism the second law could only be broken if there were many macroscopic degrees of freedom compared to microscopic ones. But dropping the assumption of two scales, we can now see that there is another possibility: using 'micro-microscopic' degrees of freedom to 'absorb' the volume in state space, while changing a large part of the kinetic energy on the microscopic level into work on the macroscopic level. Thus the macroscopic balls in the first macroscopic machine can yield their kinetic energy to a weight without breaking Hamiltonianism, because a very small amount of energy dissipates as micro-heat on the

atomic scale. And the molecular gas in the tinyon machine can be made to give up its heat in order to do work because the tinyons can use a tiny amount of energy in order to expand their volume in state space. (Please remember that all talk about volume in state space is actually disguised talk about ensembles and time evolutions of these.)

Why does the second law hold in our universe on the scale where 'temperature' refers to the kinetic energy of atoms and molecules? In chapter 3 we concluded that a necessary condition was the non-existence of vast amounts of macroscopic objects which were not made up of microscopic constituents. But this condition is not sufficient, for the same holds for macro-macroscopic objects, and yet the second law holds no longer once we scale it upwards. We can now add another necessary condition: the non-existence of tinyons, that is of tiny particles, much smaller than atoms, which can nevertheless be used to build tools and instruments to be used by Maxwell's Demon.

4.2.3 Cyclicity revisited

We are left with one question: there was dissipation in our tinyon machine and in our working macroscopic machine. In each case, a tiny amount of the transformed kinetic energy is not changed into work, but is transferred to the level of the smallest particles. As the Demon continues its operation, more and more energy will be transferred to this level – even though it need be only a very small fraction of the energy extracted as work. Does this break the cyclicity condition?

After much discussion, we decided at the end of chapter 2 that the cyclicity condition ought to be understood as the requirement that Maxwell's Demon operate "without endangering its own continued operation". Whether the tinyon machine does or does not break this requirement depends on further assumptions about the universe. If the tinyons reach a temperature (defined on their scale) above which their structures fall apart – like the melting of metal on the atomic scale – the machine will stop functioning. If tinyonic structures never fall apart, the machine will stop functioning once the kinetic energy of a tinyon is on par with that of an atom – once that happens, the trapdoor and all similar devices will stop functioning for familiar reasons. But such processes only have to take place if the amount of tinyons in the relevant part of the universe is smaller than or of roughly the same size as the amount of atoms. If there are many more tinyons than atoms, their kinetic energy never has to rise to a temperature which is too high for the machine to function. If there are enough tinyons, cyclicity as we defined it is not broken.

I already pointed out in chapter 2 that my version of the cyclicity condition is far from perfect, although it is the best I can think of. I stress at this point that adopting another version of this condition may disqualify Demons like the tinyon machine; it does not, for instance, obey the condition of strict cyclicity. To me, this disqualification seems a major disadvantage of such versions of the cyclicity condition. If we had tinyons, enough tinyons, we could run ships on the heat of the sea for as long as we would like. This is exactly the sort of thing the second law forbids. If we had tinyons, we could break the second law both in letter and in spirit. That ought to be captured in our definition of Maxwell's Demon.

4.2.4 Other measures of entropy

We have shown how the second law can be broken, at least in the sense that anti-entropic phenomena take place, if there exists a class of particles we called tinyons. An intriguing question is whether this result can be replicated using any of the quantitative measures of entropy discussed in chapter 2.

The Boltzmann entropy is lowered whenever the system enters a macrostate with a smaller volume in state space. These macrostates are defined by specific values for macroscopical quantities as well as thermodynamical quantities such as pressure and temperature – temperature on the atomic scale, that is. A first question we have to ask is whether changes in the tinyon-temperature are reflected by a change of macrostate. This seems mostly a matter of definition, so let us look at both ways of deciding the issue. If temperature changes on the tinyon scale do not affect the macrostate of the system, the possible states of the tinyons form merely a constant factor of multiplication in calculations of a macrostate's volume in state space. This means we can simply neglect tinyonic changes, and focus exclusively on the volume in state space that macrostates have due to the temperatures and pressures on the atomic scale. In this case, the tinyon machine of section 4.1.4 evidently lowers the Boltzmann entropy: it lowers the temperature of the gas without effecting any balancing processes, and thus the systems ends up in a much smaller macrostate. So the Boltzmann entropy rightfully indicates that anti-entropic phenomena are taking place. If we take the other side on the issue of tinyonic relevance, and claim that temperature differences on the tinyonic scale lead to different macrostates, we reach a different conclusion. For in that case, we move from a macrostate with relatively little volume in state space due to tinyonic degrees of freedom and relatively much volume in state space due to atomic degrees of freedom, to a macrostate where this is reversed. Indeed, considerations of Hamiltonianism will quickly persuade us that the volume of these macrostates will often be equal – hence, no change in the Boltzmann entropy has taken place. In this case, Boltzmann entropy does not indicate that anti-entropic phenomena have taken place.

It would stand to reason to adopt the first of the two choices and move on with the Boltzmannian project, were it not for a nagging doubt. The macrostates are supposed to capture all that is macroscopically detectable. But tinyonic temperature may be macroscopically detectable, depending on the exact physics of the tinyons. Suppose tinyonic structures expand when they become hotter, just like atomic structures. If we built a macroscopically large rod of tinyons, this would be an apparatus that could accurately measure tinyonic temperature. As I said before, in subsection 2.2.2, there is a lot of work to do before the Boltzmann entropy can fulfill the same role as pre-defined anti-entropic phenomena.

The fine-grained Gibbs entropy is now imbued with a truly tantalising quality. Of course, if the ensemble $\rho(\vec{x})$ is defined on a state space spanned by both atomic and tinyonic degrees of freedom, the fine-grained Gibbs entropy is constant; being a mathematical truth, Liouville's theorem is not easily broken. But if we cut away all tinyonic degrees of freedom and let $\rho(\vec{x})$ range only over the atomic ones, the fine-grained Gibbs's entropy *can* change and indeed is lowered whenever anti-entropic phenomena take place. For it now measures the 'multiplicity' divided among the atomic degrees of freedom only, and this is lowered whenever heat is converted into work (or anything else, for that mat-

ter). The prize is of course that $\rho(\vec{x})$ is no longer a complete description of an ensemble. Does this mean we now have a quantitative measure of entropic phenomena? No. No law forbids atomic heat to be changed into tinyonic heat, not even the second; this is not an anti-entropic phenomenon, even though it does decrease $\rho(\vec{x})$. But the present discussion does show us that the fine-grained Gibbs entropies on state spaces which are spanned by the degrees of freedom of single scales, might be fair candidates for indicating when entropic and when anti-entropic phenomena take place. Introducing matters of scale into the fine-grained Gibbs entropy in this manner might make it relevant to the topic of Maxwell's Demon; as is to be expected, for a topic so closely linked to matters of scale. I wish to indicate how the fine-grained Gibbs entropy could thus be used to formulate a generalised second law in appendix A. In any case, the normal, standard fine-grained Gibbs entropy defined on the whole of state space remains as useless as it ever was for demonic purposes.

In subsection 2.4.2 I promised to ignore the coarse-grained Gibbs entropy. It is a promise I will keep, but for saying here that the argument advanced against it in that subsection are still in force, and that most interesting things which can be said about it have already been said in this subsection about either the Boltzmann entropy or the fine-grained Gibbs entropy.

Neither Boltzmann nor Gibbs entropy is as such unfit to capture the matters of scale discussed in this chapter, but in both cases considerations are needed which have not hitherto been made in the general literature. In Boltzmann's case, one has to say more about how different scales influence the partition of state space into macrostates. In Gibbs' case, detailed considerations of all scales have to be made.

4.3 Conclusion: the contingent law

It is now time to summarise our findings and conclude this first part of the thesis. We tried to find out whether the statistical second law, familiar from statistical physics, was a necessary consequence of classical mechanics; and if not, what requirements we were to add in order to derive the second law. We have concluded that the second law does not follow from classical mechanics; that Hamiltonianism may provide a framework for discussing the law, but does not in any way prove its validity. In standard discussions of the second law, the world is divided into microscopic and macroscopic objects, and heat and work are defined along the lines of this dichotomy; but classical mechanics introduces no asymmetry between these two realms. Without such an asymmetry, there can be no second law. It was shown that if there were many more macroscopic than microscopic objects, the second law would not hold.

Once we left the above-mentioned dichotomy, it became clear that the second law could be formulated at any of a vast number of scales: heat could be identified not only with the movement of molecules, but also with that of macroscopic balls, stars, or tinyons. Surprisingly, the scaled up versions of the second law do not hold; and consideration of a concrete pressure demon but also arguments concerning volume in state space made it clear what is different about the scale of atoms and molecules: there is no lower scale beneath it, there are no tinyons which can be used to build a Maxwell's Demon on the atomic scale. If there were enough tinyons, we could break the second law as much

as we wished. There might be a new 'second law' looming at the horizon: one which said that the kinetic energy of tinyons could never be used to do work. And this law would hold, unless there were even smaller particles... and so on. But for the second law to hold at any level, a contingent fact about lower levels is needed; and if there were an infinity of descending scales, the second law would not hold at all, in any guise.

We can now see what kind of facts the second law is contingent on. Even within the framework of Hamiltonian physics, both the existence of and an asymmetry between different scales has to be postulated. The relevant asymmetry appears to be the fact that the atomic scale is the smallest relevant scale in our universe – nucleons and such being irrelevant because one can only build atoms from them, and thus not make any tools of them that can be manipulated but through their constituting atoms – and that the number of degrees of freedom on this scale vastly outnumbers that on higher scales. We recall Maxwell's words from section 1.3.2:

> [I]f the heat is the motion of finite portions of matter and if we can apply tools to such portions of matter so as to deal with them separately, then we can take advantage of the different motion of different proportions to restore a uniform hot system to unequal temperatures or to motions of large masses.

He can be interpreted as being right on track: if only we had tools the size of atoms but made of much smaller particles, we could create machines like the pressure demon. But we haven't and we can't, not being clever enough. Hopefully, the preceding chapters have shown what this cleverness would have to consist in, and have provided Maxwell's words with a sound backing based on the presupposed Hamiltonian nature of the world.

These results furnish us with a point of view from which we can judge reasons advanced by the exorcists for the non-existence of Maxwell's Demon. Is it true that the demon must do measurements, and that every measurement increases entropy? Is it – more relevant to recent discussions – true that the demon must erase its memory, and that this process of erasure must induce an entropy increase? That there is a deep and significant connection between entropy and information, and that understanding this connection is the key to understanding Maxwell's Demon? These questions will be tackled in the second part of the thesis.

Interlude: An imaginary history

The twentieth century has been an eventful epoch for Maxwell's Demon. Started in the second half of the nineteenth as a thought experiment by James Clerk Maxwell to elucidate the second law, before long it was perceived as a threat to a beloved bulwark of physical truth. The second law had to be protected against the malicious attempts of the infernal rascal to violate it; the lofty foundations of thermodynamics had to be secured once and for all by providing a definite and unassailable proof of the demon's non-existence. Thus began the exorcist tradition.

The history of exorcism can be divided into three main phases, in which thermal fluctuations, measurement and erasure of information were the consecutive notions thought to contain the key to any successful exorcism. The protagonist exorcists of every phase naturally saw the earlier attempts as well-meant but insufficient, and their own pet idea as the final piece of the puzzle, which at last made everything fit together and had the power to banish the demon for all eternity. Because the erasure-type exorcism is nowadays the predominant position among Maxwell's Demon scholars, the tale which displays the history of exorcism as one of gradual improvement is nowadays somewhat of an orthodoxy. It is told with zeal by Leff and Rex (2003, [20]), and by Charles Bennett (1988, [5], and very funnily in 1998, [5], figure 3). We will now present this exorcist version of the history of Maxwell's Demon, both as an introduction to the different kinds of exorcism and as a prelude to the different history which will unfold in part II – a history in which there is certainly no constant improvement detectable.

For some time after its creation by Maxwell, there were no attempts to exorcise Maxwell's Demon. But then Brownian motion of mesoscopic particles was observed and recognised as violations of the second law. This led people to wonder whether these phenomena might not be used to convert heat to work. Maxwell's Demon suddenly didn't seem as impossible as before, and attempts were made both to construct and to exorcise it. A large variety of intricate devices which could use the motions of individual molecules to produce macroscopic work were thought out, critically examined, and shown to be defective. Thus was the *first stage* of exorcism. Smoluchowski argued in his important papers Smoluchowski 1912 ([30]) and 1914 ([31]) that all machines which try to transform heat into work by exploiting thermal fluctuations, must themselves also be defeated by thermal fluctuations. The paradigmatic example is the one-way valve, in which the trapdoor starts jumping up and down at random and

stops functioning due to the thermal fluctuations which it was meant to exploit. This stage of exorcism gave us several insightful descriptions of the failure of specific systems, but unfortunately no general proof that no Maxwell's Demon is possible.

The *second stage* of exorcism started in 1929 with the publication of Szilard 1929 ([32]). In this paper, Leo Szilard claims that Maxwell's Demon is really defeated by the fact that it has to do measurements, and these measurements have an irreducible entropy cost which exactly nullifies any entropy-decreasing actions the demon could possibly undertake. Putting it more succinctly: the demon must acquire information in order to operate, and information acquisition generates entropy. There was, however, nothing really resembling a proof in Szilard's paper. Later, Brillouin 1951 ([7]) furnished such a proof for Maxwell's original demon, equipped with a torch, trying to sort slow and fast molecules. He showed that, indeed, doing measurements would cost the demon more than it could gain; according to Brillouin himself, this was a consequence of quantum mechanics, but later authors casted doubt on this idea. At this stage of exorcism, we were already much farther than Maxwell, who had not considered the measurements his demon would have to do at all, but not yet far enough: there was no general proof that information acquisition implies entropy generation. And there would not be one, because it is simply false: one can do measurements without any entropic cost.

The last step towards clearing up the mystery of the demon was taken by Rolf Landauer, in his Landauer 1961 ([17]). Here he showed that the only information processes which required entropy generation were logically irreversible ones. Now all operations can be made logically reversible, except for one class of them: information erasures. Information erasure always implies entropy generation. And this is the final clue about Maxwell's Demon: because it has to operate in a cycle, it must erase its own memory of what it has measured at the end of a cycle. And this erasure will always, this can easily be shown, generate at least as much entropy as the tiny hellion could have gained. Thus, recognising this cost of information erasure defeats Maxwell's Demon. Better still, we can prove Landauer's Principle – the claim that erasing one bit of information generates $k \ln 2$ units of entropy – and use it to get a completely general exorcism, no longer limited to a few special examples. Thus, the *third stage* of exorcism has been a more or less complete success.

This is the imaginary history which exorcists have told us. In the second part of the thesis, I will argue that the real history of Maxwell's Demon is fundamentally different.

Part II

Tales of the exorcists

Chapter 5

Doomed by fluctuations

5.1 Aims and claims

Part II of this thesis has two main aims. The first is to show that the exorcists, who claim to have banished the demon on necessary grounds, have done no such thing. Such a demonstration is necessary to uphold the conclusions of part I. The second, more important, aim is to present and refute several claims that the exorcists have made about Maxwell's demon. Foremost among these is the idea that somehow, entropy and information are inextricably linked. I will presently give an overview of what I will do in the next three chapters.

In chapter 5, we will look at the first stage of exorcism, in which the demon was thought to be defeated by thermal fluctuations. I will claim that this idea contains much truth, but that any demonstration of this principle needs an assumption (the assumption that all parts of the contraption are to be described by the canonical distribution function) which is comparable in strength to the facts identified in part I as those on which the second law is contingent. Thus the exorcism, though sound, does not succeed in defeating all Maxwell's Demons. In addition, there is a gap in its argumentation concerning the doing of measurements.

Measurements therefore take centre stage in chapter 6. Here, two claims of the exorcists are examined and found wanting. The first is that the measurement-account is a completely new stage of exorcism. Most of its sound physical core is reducible, so I argue, to the fluctuations-account. The only difference is that some important but quite peripheral considerations of measuring have been added to the basic ideas of Smoluchowski and his contemporaries. Their second claim is that there is a deep connection between entropy and information acquisition. I counter Brillouin's arguments to this effect, and propose that adding talk about information to the fluctuation/measurement-account of the demon adds not one iota to our understanding.

Then, in chapter 7, information returns, this time in the guise of Landauer's Principle and the claim that considering the effects of information erasure is the key to defeating Maxwell's demon. I argue that as far as the Landauer-Bennett scheme is successful, it reduces to either the claim that all systems must be described by a canonical distribution function, or to a version of the State Space Contraction argument from chapter 3. Neither of these two has

anything to do with information. The idea that entropy and information are closely linked is not defendable, I claim, and Landauer's Principle is not the butter and bread of exorcism.

Ironically, the history which emerges from these polemics is one of gradual decline rather than of constant improvement.

5.2 Doomed by fluctuations

5.2.1 The one-way valve

Before commenting on the scheme as a whole, we will look at some examples of exorcisms within the fluctuation tradition. Smoluchowski's one-way valve has already been described in section 4.1.1. Let us take a quick look at the basic idea behind this device. In order to be a *one*-way valve, the machine must ensure that (almost) no molecules from B can enter A, whereas those from A *can* enter B. The hole in the container wall, then, has to be closed most of the time – so that no molecules from B will pass – and must be opened whenever a molecule in A tries to pass. The molecule in A has an average kinetic energy of $\frac{1}{2}kT$ in a given direction, so the trapdoor must be sensitive to energies that high, and preferably lower energies as well. But the trapdoor has the same temperature as the molecules; it is, after all, in contact with them. That means it has fluctuations of kinetic energy of the same size as the molecules; that is, on the size of kT. The trapdoor must be sensitive to energies of order kT, and it itself is plagued by fluctuations of order kT. So it is sensitive to random fluctuations, and there will be no correlation between the openings of the trapdoor and the arrival of molecules from A. This reasoning not only works for pressure demons, it is just as valid for temperature demons. Such a demon must be a sorting device, sorting hot and cold molecules. But in order to sort it must be sensitive to kinetic energy differences of order kT, and the story repeats itself.

But couldn't we use a trapdoor, or a sorting device, which is cooled? If it were colder than the gas, it would not have to be defeated by fluctuations. This is true, and it is very important to realise that a trapdoor could work that way – but only at a price. During its contact with the gas and the rest of its environment, the cold parts of the machine would invariably heat up. There is a transfer of heat from a hotter to a colder body, from the gas to the device. This is an entropy increasing process. The demon uses a temperature difference, that between the gas and the device, in order to do work; this is what we called 'cheating' in Part I. The recognition of this possibility of cheating is central to the next example.

5.2.2 Ratchet and pawl

Perhaps the best known example of a system supposed to violate even the statistical second law of thermodynamics is the ratchet and pawl device made popular by Feynman in his famous lectures on physics.[1] Imagine a box A full of gas at a certain temperature T_A, and another box, B, at temperature T_B. In box A there is a set of vanes fastened onto one end of an axle in such a way

[1] Feynman et al., 1963 [12], chapter 46.

that pressure fluctuations can move the vanes and turn the axle. Assume that we make this device so small and light that it can actually be set into motion by the kind of pressure fluctuations regularly occurring in the container. What is going to happen is that the vanes and the axle will turn back and forth, now going one way, then the other, without any average rotation to be expected. We cannot, therefore, lift a weight through the turning of the axle. In order to do this, we would have to restrict the turning to one direction. This is were box B comes in: the axle goes out of box A and into box B, where a wheel which can turn only one way is fastened onto it: the ratchet and pawl. Now whenever the axle turns one way, the wheel will turn with it, lifting the pawl until it reaches the top of a tooth and snaps back into place, preventing the wheel from turning the other way.

Let us consider the machine in a little more detail. The pawl, in order to be able to snap back whenever it has passed a tooth of the wheel, must be fitted with a spring which pushes against it. We will call the energy needed to raise the pawl to the top of a tooth against the force of the spring ϵ. Furthermore, this energy must be dissipated when the pawl snaps back. If it weren't, the pawl would collide elastically with the wheel and rebound, allowing the wheel to reverse its motion. The energy ϵ has to be dissipated when the pawl and the wheel collide; therefore, the pawl, the wheel and the gas in box B will heat up. Still, this does not refute the ratchet and pawl device as a successful Maxwell's Demon: whatever the values of T_B and T_A, our current analysis would predict that T_B would rise and T_A fall. For $T_B < T_A$, this would be an example of a cheating demon: heat is transported from a hotter to a colder body, which implies raising the entropy. For $T_B > T_A$, this would mean that heat is transported from a cold to a hot object, an anti-entropic effect.

But alas, it is not to be. The probability of a statistical mechanical system with temperature T being in a state with energy E_1 is proportional to $e^{-E_1/kT}$ – this is just the well-known canonical distribution function. Thus, the probability of the system gaining enough energy through pressure fluctuations in container A to turn the pawl over a tooth is proportional to $e^{-\epsilon/kT_A}$. And the probability of the pawl gaining enough energy to jump up by itself, allowing the wheel to turn the wrong way, is proportional (with the same constant of proportionality) to $e^{-\epsilon/kT_B}$. Evidently, for the device to work the first event should happen more often than the second; mathematically $e^{-\epsilon/kT_A} > e^{-\epsilon/kT_B}$. But this implies: $T_A > T_B$. So the device can only work if the temperature in container A is higher than that in container B; but as we have seen, then the device is just a machine using a temperature difference to do work, which is not a violation of the statistical second law. To violate that law, we need $T_A \leqslant T_B$. But in that case the pawl, because of fluctuations in container B, is bouncing up and down so frequently that it no longer fulfills its appointed role.

5.2.3 Three hot bodies

A third and last example – we could go on almost indefinitely if we wanted to – is discussed by Smoluchowski 1914 ([31], p. 118-119). When two bodies are brought into thermal contact, they will exchange heat. Thermodynamics tells us that the hotter body will cool off while the colder body heats up, until both have the same temperature and equilibrium is attained. In statistical physics, however, fluctuations will arise in the temperatures; big fluctuations

are more rare than small ones, but if we wait long enough even moderately sized temperature differences between the two bodies will arise.

Smoluchowski proposes a machine which will exploit this fact to reliably create an arbitrary temperature difference between objects originally at the same temperature. Imagine three bodies, A, B and C, all at temperature T. The machine will use B to transport heat from A to C. First, B is brought into contact with A, until a lucky fluctuation occurs in which the temperature of B rises ΔT above that of C. At this moment we break the contact between A and B, and bring B into contact with C. Now we wait until a fluctuation has arisen in which B cools below the temperature of A. We disconnect B and C, and bring B once again into contact with A, to start another cycle. If successful, every cycle will transport some heat from A to C, never the other way around. Thus, we will have a successful Maxwell's Demon, which invariably takes two bodies of equal temperatures, A and C, then transports heat from one to the other.

But one step of the operation of this automatic machine has not yet been specified: measuring the temperature of B in order to decide when it should move to the other body. We'll equip B with a thermometer and a mechanical device coupled to it, such that it moves B to A whenever the temperature falls below one threshold, and moves B to C whenever the temperature rises above another (higher) value. The reader will have anticipated the problem that Smoluchowski now raises: this auxiliary device becomes a victim of fluctuations. Smoluchowski stresses that these fluctuations in the auxiliary device are not correlated with the thermal fluctuations between the two bodies in contact. Therefore, B will start moving regardless of its actual temperature, thus destroying the machine's ability to function properly.

Smoluchowski does not support his conclusion with a detailed analysis of the fluctuations arising in the thermometer and the auxiliary device, which will depend on their exact constitution. There may be problems lurking here; we will return to this in the next section, and especially in chapter 6.

5.3 Critical discussion

5.3.1 The fundamental assumption

The mechanism by which the one-way valve was defeated is easy to understand. It is in thermal contact with its environment, since molecules constantly collide with it. Because of the random motion thus imparted to it, the valve jumps back and forth, nullifying its ability to act as intended. Something similar happened to the pawl in the ratchet and pawl device. Whenever the temperature of container B is higher than that of A, the kinetic energy obtained by the pawl through thermal interactions will regularly cause it to jump up spontaneously, letting the wheel turn the wrong way. This is simply the same mechanism that was at work in the first example. But as we had two containers with potentially different temperatures here, another step was needed in the argument: a proof that if the temperature of B were lower than that of A, it would rise until equilibrium had been attained. This behaviour was ensured by an energy exchange (through mechanical means rather than direct thermal contact) that took place every time the vanes lifted the pawl.

These exorcisms are correct, given two assumptions. The first is that cooling the crucial parts of a device must always nullify its creation of work. Rhetorically, I claimed that an energy exchange from a hot to a cold body was to be considered cheating, and would disqualify a demon. But this is not quite correct. Suppose that the the cold parts of the machine are constantly cooled, and the excess energy is pumped back into the gas. This would ensure that the machine could go on operating indefinitely, if there was enough work available to keep this cooling process running. So we have a machine which on the one hand needs work for cooling, but on the other hand produces work through its operation. As yet we have not seen a proof that the work produced cannot outweigh the work needed for cooling – if it does, the machine is a working demon, fluctuations notwithstanding. I must confess that I do not have a general proof either, but the State Space Contraction argument of chapters 3 and 4 makes it quite plausible that no Maxwell's Demon can be successful, and *a forteriori* that the assumption holds, for all atom-based Maxwell's Demons with few macroscopic degrees of freedom.

The second assumption is that the valve and the pawl are indeed subject to fluctuations in kinetic energy in accord with the temperature of their surroundings. This is not proven from a detailed consideration of the microstructure of the valve or the pawl and its interactions with the environment; it is derived from the canonical distribution function, which is assumed to hold for every system including the valve and the pawl. This may seem to be a very trivial requirement: of course they too must obey the fundamental laws of statistical mechanics! But this attitude is indefensible: the assumption that every part of the demon is described by the canonical distribution function is both *essential* to the previous exorcisms, and *very significant*. It is certainly not a distribution that holds for all physical systems – consider for instance the planets of our solar system. And, more relevantly, it is not a distribution that holds for the parts of the tinyon machine of subsection 4.1.4 either. They are not described by a canonical distribution function using the T of the atoms, but – probably – by a canonical distribution function using a different 'temperature', τ say, which is only defined on the tinyonic scale. The tinyonic trapdoor does not fluctuate with the same kinetic energy fluctuations as the molecules it is meant to stop or let pass do, because it is not described with the same canonical distribution function. It does not even *have* a temperature akin ot that of the molecules, just as the atoms of the working macroscopic machine did not have a temperature akin to that of the balls which bumped around in it, and the particles in a star have nothing whatsoever to do with the 'temperature' of the star cluster. What the assumption of the canonical distribution governing all parts of the machine actually says is that the temperature it talks about describes everything – that, in other words, the scale on which this temperature is defined is the definite scale, the only important scale, the point where our investigation can stop. Thus, the tinyon machine is excluded by assumption, not by argument. I will call the assumption that all physical systems must be described as having a temperature which can easily be compared to all other temperatures – the assumption, in other words, of only one relevant scale of thermal phenomena – the **fundamental assumption**. The fluctuation school of exorcism may teach us very well why certain actual machines fail to work, but it does not prove – I repeat, it does *not* prove – that Maxwell's Demon is forbidden by the laws of nature. It needs the very strong fundamental assumption to get off the ground.

Furthermore, the two assumptions identified together imply that a working demon must be at the same temperature as its surroundings – because cooling is not efficient, and the normal temperature scale is applicable to it. Within the framework of thermal physics, this can be interpreted as the claim that if a working demon is in an environment with temperature T, then it and all it surroundings must be describable by a canonical distribution function for temperature T. I will call this the **extended fundamental assumption**. It will return in the next chapter.

Using either of the fundamental assumptions might be enough to disprove the possibility of Maxwell's Demon. Such a proof would then be sound, in the sense that it would be a valid deductive argument, but it would not be profound, in the sense that its assumptions are so strong that we have not really learnt whether Maxwell's Demon can or cannot exist – we would still need to know whether the assumptions do or do not hold. John Earman and John Norton have claimed that all purported proofs of the demon's nonexistence are either unsound (not valid deductive arguments) or not profound (based on assumptions too strong to be believed by the demon's proponents). We will return to their claim in subsection 7.1.3.

5.3.2 The issue of measurement

The third system, Smoluchowski's three-body machine, appears to be based on the same principles as the previous examples: defeat by unintended fluctuations. And the apparent dissimilarity that it uses temperature fluctuations to perform its task instead of pressure fluctuations is of little importance. Nevertheless, there *is* a remarkable difference. In the first example, the one-way valve was in thermal contact with the gas it had to sort. Because the molecules of the gas imparted energy to it, it heated up and stopped functioning. In the second example, the ratchet and pawl where in direct mechanical contact with the vanes, having the responsibility to prevent their turning in the wrong direction. And the vanes, in return, had to impart an amount of energy ϵ to the pawl whenever they wished to turn. But in the third case, no such exchange of energy seems to be necessary. The mechanical moving of B can be done by an adiabatically isolated machine, which would not receive any energy from B. That still leaves the temperature measurement; does this involve an exchange of energy that would somehow destroy the ability of the machine to work? Wouldn't it be possible simply to *look* at system B, without heating up? Or consider Maxwell's original demon. It does not have to fail like the one-way valve did, since it never has to come into contact with the molecules. If it just looks at them, and opens its shutter accordingly, why would it need to become hot like the one-way valve and the ratchet and pawl? Obviously, what is needed here is an analysis of the Demon's measurements: do they in some way or another have the same effect as the energy exchanges taking place in the one-way valve and the ratchet and pawl? This analysis will be carried out in chapter 6.

Chapter 6

The details of measurement

6.1 The cost of measurement

The previous chapter ended with the claim that an analysis of measurement is needed if we wish to complete the fluctuation account of exorcism. This project has indeed been taken up by several people, but it led to the much more far-reaching conclusions that all measurements had an entropy cost, and that this was the key to understanding the demon. Going yet further, it was argued that entropy and information acquisition where very closely linked, and one needed to pay attention to information

6.1.1 Szilard's engine

The first attempt to connect measuring and entropy was made by Leo Szilard in Szilard 1929 [32], one of the most influential papers in the history of Maxwell's Demon. Instead of discussing his original machine, I'll describe the simplified 'Szilard engine' which is now almost exclusively used in the literature. Imagine a cylinder with a volume V_1, which contains exactly one molecule of gas. This cylinder is in contact with a heat bath of temperature T_1, which ensures that the molecule has an average kinetic energy corresponding with this temperature. An ensemble of these systems is described by the canonical distribution function $e^{-\epsilon/kT_1}$. The demon now places a partition in the middle of the cylinder, dividing it into two equal chambers with volume $V_1/2$. It takes a look at the molecule, finding out in which half of the cylinder it has been trapped, and places a piston in the other half, with a small weight attached to it with a string. After this procedure, the demon removes the partition and the molecule of gas will start doing work on the piston until it has been moved to one side entirely and the molecule can once again occupy the entire volume. The weight has been raised making use of the thermal energy of the heat bath. (Obviously, this scheme only works if the weight is so light that the gas pressure is high enough to overcome the pull of gravity on the weight.) Now the piston is removed, and the system has gone through a complete cycle, ready to begin anew.

There are three ways of attacking this demon. The first is to deny that the system has gone through a cycle at the end of the described operations. The demon, one can claim, has memorised the result of his measurement; this result must be erased before the cycle has been completed. And memory erasure,

the claim goes, must always lead to an entropy increase. Hence, a careful analysis of the demon's mnemonic faculties will show that the engine does not work as promised. This very popular argument will be discussed in chapter 7. The other two ways to attack the demon focus on the process of measurement. They involve the claim that during its measurement something has happened which nullifies its successes (the second way) or stops it from operating (the third way). The latter option tries to show that measuring leads to the kind of fluctuations in the measurer which defeated the valve and the ratchet and pawl device; I will discuss this possibility in subsection 6.1.4. The former option, which encompasses the idea that measuring increases entropy by as much as the machine's cycle can lower it, is that which Szilard chose; he wrote:

> One may reasonably assume that a measurement procedure is fundamentally associated with a certain definite average entropy production, and that this restores concordance with the second law. The amount of entropy generated by the measurement may, of course, always be greater than this fundamental amount, but not smaller.[1]

There are two things we would like to know. One is what Szilard means with 'entropy', the other how doing a measurement generates this entropy.

Actually, for entropy we are just going to use the thermodynamical definition. Let us calculate the amount of work done on the weight by the molecule as it expands. The formula that gives us the work done by an ideal gas expanding from volume V_1 to V_2 at pressure P is:

$$W = \int_{V_1}^{V_2} P dV. \tag{6.1}$$

Using the ideal gas law

$$PV = NkT \tag{6.2}$$

and the fact that $N = 1$, we arrive at the following result:

$$\begin{aligned} W &= \int_{V_1/2}^{V_1} \frac{kT_1}{V} dV && (6.3)\\ &= kT_1 \ln(\frac{V_1}{V_1/2}) && (6.4)\\ &= kT_1 \ln 2. && (6.5) \end{aligned}$$

This energy comes from the heat bath in the form of heat, so the amount of heat Q extracted from it is $-kT_1 \ln 2$. Using formula 1 from the Prologue:

$$\int_A^B \frac{đQ}{T} = S_B - S_A, \tag{6.6}$$

it is easy to show that the entropy of the heat bath has been changed by an amount

$$\Delta S = -k \ln 2. \tag{6.7}$$

The use of the thermodynamic entropy in this derivation is unobjectionable: we are simply talking about objects exchanging heat and performing work on

[1] Szilard, 1929 [32]; from [19], p. 127.

each other, which is what thermodynamics is all about. So presumably when Szilard claims that the measurement must – on average – produce entropy, he is still talking about thermodynamic entropy. Somewhere in the measuring process, heat must flow from a hotter to a colder body, or a weight must be lowered to produce heat, or some other entropy-increasing thermodynamic process must take place. Perhaps something similar to what happened in the ratchet and pawl device is going on, where the decrease of entropy in the box with the vanes was compensated by an increase of entropy in the box with the pawl – the former cooled, but the latter heated up, and the device only worked if the latter was colder than the former. In order for Szilard's entropy increase to be possible at all, our demon and the devices he uses must be ordinary thermodynamical objects with well-defined temperatures between which heat exchange and similar operations can take place. This is, in effect, the fundamental assumption I pointed out in subsection 5.3.1.

Anyway, we are by now very interested in the way in which all of this works in actual measuring processes. Unfortunately, Szilard does not really analyse these in terms of thermodynamical operations; he merely concludes that an entropy increase of $-k \ln 2$ must be associated with a measurement, because otherwise the second law doesn't hold. As an exorcism his article is a failure: it assumes the validity of the second law. As a dissection, it does not fare much better: we still don't know which facts about measurements save the second law. But it nevertheless proved to be a very good starting point for further research into the connections between entropy, information and measurement. The first major step forward was made when Brillouin and other researchers gave detailed analyses of concrete measurements. We turn to that subject now.

6.1.2 Brillouin's torch

In 1951, more than two decades after Szilard wrote his paper, Léon Brillouin published[2] an analysis of Maxwell's Demon which focussed on the way in which it was to do measurements. Brillouin took literally the idea that the demon has to 'see' the molecules it is working with, and calculated the entropy cost associated with their visual detection. Recall the original thought experiment by Maxwell, in which a demon functions as a doorkeeper which lets some molecules pass and stops others. If it has no tools to help it, Brillouin remarked, it will not be able to see the molecules. The container and the gas within it are at a uniform temperature T, and thus all of the space surrounding the demon will be filled with a homogeneous blackbody radiation. There is no difference in photon density between a container full of molecules and one which is empty, let alone that individual molecules can be seen. Therefore, the demon cannot operate unless it is equipped with a torch which emits photons of a wavelength considerably different from the most common wavelengths as specified by Wien's law:

$$\lambda_{max} = \frac{0.002898}{T}, \tag{6.8}$$

where T is in Kelvin, and λ in meters. Brillouin provides the demon with a charged battery and an electric bulb, which together act as a source of blackbody radiation at temperature T_1. The gas is at temperature T_0, and to obtain

[2] Brillouin, 1951 [7].

light which can be distinguished from the background radiation, we must have $T_1 \gg T_0$. The average energy $h\nu_1$ of a photon emitted by the bulb is of the order of kT_1, so:

$$h\nu_1 \gg kT_0. \tag{6.9}$$

Now assume that during the demon's operation the battery yields an energy E and no entropy to the filament. In turn, this radiates a total energy E, therefore decreasing its entropy by $S_f = E/T_1$. If there were no demon making use of this light, it would be absorbed by the gas, giving an entropy increase of $S = E/T_0 > S_f$. Thus, the total entropy would increase.

But this is where Maxwell's Demon comes in. To detect a molecule, at least one quantum of light must scatter on it and enter the Demon's eye, which has a temperature of T_0. This represents an increase of entropy

$$S_d = h\nu_1/T_0 = kb, \tag{6.10}$$

where

$$b = h\nu_1/kT_0 \gg 1. \tag{6.11}$$

So for every seen molecule, the demon must pay a price of kb units of entropy. Now assume that Maxwell's Demon, operating on two containers A and B, has already succeeded in creating a temperature difference ΔT:

$$T_B = T + 0.5\Delta T \tag{6.12}$$

$$T_A = T - 0.5\Delta T. \tag{6.13}$$

The demon must observe one molecule in both of the containers, allowing a fast one from A to enter B, and a slow one from B to enter A. The entropy cost of these two detections is $2kb$. Now assume that the molecule from A has a kinetic energy $1.5kT(1+\epsilon_1)$, and the molecule from B a kinetic energy $1.5kT(1-\epsilon_2)$. This results in an energy, and hence a heat, transfer of

$$Q = 1.5kT(\epsilon_1 + \epsilon_2) \tag{6.14}$$

from A to B, which corresponds to an entropy decrease of

$$\Delta S_i = Q(\frac{1}{T_B} - \frac{1}{T_A}) \tag{6.15}$$

$$= -Q\frac{\Delta T}{T^2} \tag{6.16}$$

$$= -1.5k(\epsilon_1 + \epsilon_2)\frac{\Delta T}{T}. \tag{6.17}$$

Since the demon cannot choose before measurement which molecules he is going to use, the quantity $\epsilon_1 + \epsilon_2$, representing the deviation of the molecules from the mean kinetic energy in the gasses, will generally be rather small, reaching a few units only exceptionally. Furthermore, the coefficient $\Delta T/T$ will be much smaller than unity. Hence,

$$\Delta S_i = -1.5k\eta, \qquad \eta \ll 1. \tag{6.18}$$

Therefore, the total entropy change of the gas is:

$$\Delta S_d + \Delta S_i = k(2b - 1.5\eta) > 0. \tag{6.19}$$

The thermodynamic entropy, which we have been using throughout, has increased; so no anti-entropic phenomena have taken place. Maxwell's Demon has been defeated by the measurements it had to make.

6.1.3 Critique of Brillouin's argument

The argument given in the previous subsection is intriguing and suggestive; but is it correct? The physics is relatively clear, but I perceive several questions left unanswered by Brillouin's account. First, however, let me discuss two possible unclarities.

> **The role of the battery has not been discussed properly. Does it not constitute an external source of power of the kind we forbade in section 2.5?**

If the demon has a battery which it uses to heat up a filament, it has an external source of power which could have been used to do work directly. One might think that this is an immediate refutation of the demon (remember section 2.5), but that is too hasty. If the amount of work which could have been done by the battery is lower than the amount of work that has been extracted from the gasses by the demon, one can recharge the battery and still have some work remaining. Therefore, using a battery is not something a demon may not do; but recharging the battery should be part of its cycle. We now turn to the second possible unclarity.

> **Has an entropy increase associated with the heating of the filament not been ignored?**

If we start out with a cold filament and a full battery, the heating of the filament indeed constitutes an entropy increase, since work is transformed into heat. But let us consider the very similar case in which the filament starts out at temperature $T_1 \gg T_0$, and the task of the battery is to add as much energy to the filament as it sends away in the form of photons. This raises the entropy, which seems to be ignored by Brillouin; but actually it is not. Adding energy $h\nu_1$ to the filament at temperature T_1 increases the entropy by $h\nu_1/T_1$. Sending a photon with energy $h\nu_1$ from the filament to the demon corresponds with an entropy increase of

$$\Delta S_{f\to d} = \frac{h\nu_1}{T_0} - \frac{h\nu_1}{T_1}. \tag{6.20}$$

So the total entropy increase of the process is merely the sum of these two, $h\nu_1/T_0$ – which is exactly the value used by Brillouin when he discusses the entropy cost of detecting one photon. This should satisfy us that nothing is wrong with the way Brillouin uses a battery in his argument.

Now that these unclarities have been dealt with, let us ask several critical questions that may threaten our belief in the validity of Brillouin's argument.

> **Can't the demon use light of a lower, instead of a higher, frequency than that at which the black-body radiation peaks?**

In order to be able to distinguish the photons reflected by the molecules from those of the background radiation, they must have a different energy. But why should the demon choose to use photons of a higher, rather than of a lower frequency than that dominant in the background radiation? In that case, b would be much smaller than 1, instead of much larger, and Brillouin's argument would not work.

Leff and Rex (1990 [19]) give two arguments for preferring higher frequencies over lower frequencies. The first is that the power of a lamp is proportional to AT^4, where A is the radiating surface – this is the law of Stefan-Boltzmann. So if we choose a temperature $T_1 = 0.5T_0$ instead of $T_1 = 2T_0$, the surface of our lamp would already have to be 256 times bigger to achieve the same output. The second is that high frequency radiation leads to more pronounced diffraction effects, enlarging the resolving power of the demon.

Since resolving power does not enter Brillouin's analysis at all, it is hard to see how the second argument could save him unless it was backed up by a discussion of the demon's need to resolve – which might well turn out to be a very complicated matter. The first argument is not sufficient as it stands either – why should we care about surface area? – but points the way to something we have forgotten to take account of: the absorbtion of background radiation by the filament. Assuming that the filament is both a ideal radiator and an ideal absorber (the definition of a black body), its emitted energy is $\sigma A T_1^4$, with σ Stefan's constant, and the energy it absorbs is equal to $\sigma A T_0^4$. In that case the energy flux from the filament to its surroundings is:

$$\Delta E = \sigma A(T_1^4 - T_0^4). \tag{6.21}$$

In the case of $T_1 \gg T_0$, this is approximately equal to $\sigma A T_1^4$, and Brillouin's analysis applies: the entropy decrease which is the result from photons of the (colder) gas hitting the (hotter) filament can be neglected. But in the case of $T_1 \ll T_0$, there is a significant entropy increasing process as yet unaccounted for: the hot gas is heating the cool filament. For every photon radiated by the filament, and thus of use to the demon, many photons from the gas are absorbed by the filament. So for every measurement of a molecule, the demon must pay the cost associated with this heating of the filament, which is:

$$\Delta S_{g \to f} = \frac{nh\nu_0}{T_1} = kb, \tag{6.22}$$

where n is the amount of photons hitting the filament for every photon it absorbs and

$$b = \frac{nh\nu_0}{kT_1}. \tag{6.23}$$

Obviously, $h\nu_0 = kT_0$, so if $T_0 = T_1$ we have $b = 1$. The average energy of a photon increases linearly with the temperature of the body (Wien's law); but its emitted power increases with T^4. Thus, the number of photons emitted increases with T^3. Define $\eta = T_1/T_0$; then $n = \eta^{-3}$. So we have for $T_1 \ll T_0$ (in other words, $\eta \ll 1$):

$$\begin{aligned} b &= \frac{nh\nu_0}{kT_1} && (6.24)\\ &= \frac{nh\nu_0}{k\eta T_0} && (6.25)\\ &= n\eta^{-1} && (6.26)\\ &= \frac{1}{\eta^4} && (6.27)\\ &\gg 1. && (6.28) \end{aligned}$$

This is the same result which was reached in formula 6.11 for $T_1 \gg T_0$, and from this point on the same reasoning applies to the problem. So, even when the demon uses a black body at a much lower temperature than that of the gas, the entropy cost which must be paid per photon used defeats all his attempts.

Why must the demon rely on black-body radiation?

But a more devastating problem looms ahead. Our calculations indicate that a demon using a black body to supply it with photons cannot operate successfully. But why should it use a black body? Since the demon must be able to see the difference between its scattered photons and the photons of the background radiation, its light source must have an emission spectrum different from that of a T_0 black body. *One* way to achieve this is by taking a black body radiator at $T_1 \neq T_0$. But *another* way is to use a light source which does not emit a black-body spectrum, but perhaps a sharply peaked one. If this peak – say, an emission line – were to be at a frequency considerably less than the top of the T_0 black-body spectrum, it is hard to see how Brillouin's analysis can be used to defeat the demon. Neither his original argument that $b \gg 1$, nor the argument for the same conclusion given in the last paragraphs would be straightforwardly applicable. We can imagine a demon equipped with a small glass box full of a gas which is at the same temperature T_0 as its surroundings, but has a characteristic low-frequency emission line. It has not been shown that this 'torch' is not good enough. But unless that is shown, Brillouin's exorcism is not good enough: he has not proven that the demon cannot operate by using a suitable source of light. (However, see subsection 6.1.4, where I suggest that what is really important in the story of measurements is the fact that there must be a way to exchange energy between the demon and the system, and that the details of measurement are not all that important. If such considerations are correct, low-frequency emission lines and other special sources of light will not help the demon.)

Must de demon absorb the photons? Can't he simply reflect them?

Brillouin's argument is based on the idea that the demon absorbs the photon when he makes a measurement, and can therefore be described as a system at temperature T_0 gaining an amount of heat $h\nu_1$. It might be suggested that the demon does not have to absorb the photon: perhaps he can reflect it with a simple mirror. But this would not help, as the photon would still be absorbed by the gas or the container, resulting in exactly the same entropy increase. (Not so, of course, when the photon would be re-absorbed by the filament. But this chance is very small, as the demon – being ignorant of the next photon's direction of incidence – cannot position his mirror in the right way to accomplish this.)

Does the demon have to be at temperature T_0?

The entropy cost of the measurement is $h\nu_1/T_0$, where T_0 is assumed to be both the temperature of the gas and that of the demon. The reasoning behind this is clear: if the demon has another temperature, there will be heat exchange between it and the gas, which is an additional source of entropy. In the end,

putting the demon at a temperature $T_D > T_0$ would not help. But it should be noted that it is an essential presupposition of Brillouin that the demon is a thermodynamical system which can be assigned a definite temperature. Without this, his analysis cannot come off the ground; the thermodynamic entropy of a process can only be calculated when every system is a thermodynamic system. Brillouin needs the fundamental assumption as much as any other exorcist.

6.1.4 Relation to fluctuations

A lot has been made of the role of measurement in the Maxwell's Demon thought experiment. Brillouin's analysis has been widely hailed as a proof that the demon is actually defeated by the fact that it has to acquire information about the molecules on which it wishes to operate. Later writers such as Bennett and Leff and Rex, who disagree with Brillouin that the entropy cost is associated with the measurement, still claim that Szilard and Brillouin were the first to understand that the key to exorcising Maxwell's Demon lay somewhere in its use of information. But such a radical division between Smoluchowski on the one hand and Szilard and Brillouin on the other is a figment of the imagination. In fact, I would like to suggest that Brillouin's exorcism is closely related to that of the one-way valve and the ratchet and pawl, an exorcism in which information and measurement play no role whatsoever. I will use the idea which I pointed out as the fundamental assumption in all fluctuation-based attempts at exorcism: every system can be assigned a temperature on the same, commensurate, scale. Thus, we can claim that every systems is is described by the canonical distribution function associated with the temperature of its immediate surroundings.

Therefore, Maxwell's Demon, which is at a temperature T_0, is a victim of thermal fluctuations of order kT_0. In order to detect the photons emitted by the torch and reflected by the molecules, the demon needs a detection mechanism with two energy levels. The characteristic frequency of the photons is $h\nu_1$, so except for an arbitrary additive constant the two possible energies of the detection mechanism ought to be $E_0 = 0$ and $E_1 = h\nu_1$. There are two possible cases: $h\nu_1 \ll kT_0$, and $h\nu_1 \gtrsim kT_0$. Suppose that $h\nu_1 \gtrsim kT_0$. In order to re-use the detection device, this energy must be dissipated in the demon or its environment, both of which have a temperature T_0. This corresponds with an entropy increase of $h\nu_1/T_0$, which is more than the demon can hope to compensate for by sorting. (As has been shown by Brillouin's analysis of the average entropy benefit of sorting.) On the other hand, suppose that $h\nu_1 \ll kT_0$ – as it will be when the demon uses the low-frequency light source we devised for it in the last section. In that case, the system will constantly switch between the two possible states of the detection apparatus (and everything set in motion by it) because of thermal fluctuations. The demon's 'eye' is useless – doomed by fluctuations. It will open and close the shutter at random, and no sorting takes place.

If we assume that the demon is a thermodynamical system at the same temperature as its surroundings, its detection mechanism is subject to fluctuations. If the energy threshold for detection is too low, the detector (and the rest of the mechanism) doesn't work because of these thermal fluctuations. But if the energy threshold is too high, detection involves a heat transfer which raises the demon's entropy more than it can lower that of the gas. What we have to

keep from Brillouin's analysis is the calculation of the average entropy gained from sorting and the insight that the measurement is a relevant step in the demon's operation. But what we do not need is a detailed discussion of the actual measurement process: merely thinking about the detector and its fluctuations does the job quite nicely. This suggests that the paradigm of exorcism is still the fluctuation account of Smoluchowski, and not the measurement account of Brillouin.

And there is something else which we have learned from Brillouin. The fact that the demon must do measurements implies that it must exchange energy with its surroundings: it cannot be completely isolated. Therefore, it cannot operate at an arbitrarily low temperature where it would not experience any relevant fluctuations: because of the necessary interactions (such as the absorbtion of background radiation), the demon must operate at the temperature of its surroundings or pay an entropy cost to remain cool – and this entropy cost would be too high for it to continue functioning. This is a valuable insight: even a demon which does not make physical contact with the gas exchanges energy with it because it has to do measurements. Brillouin's argument could be reinterpreted as an attempt to show that if the *fundamental assumption* of subsection 5.3.1 holds for a demon and every part of its surrounding, the *extended fundamental assumption* must hold too.

The conclusion of this section is that Brillouin's account of measurements is not an entirely new way of approaching the demon, but a valuable addition to the older account in two ways. First, it points to the fact that an accurate measurement made by a demon at the same temperature as its surroundings involves an entropy cost. Second, it stresses that every demon must have interactions with its surroundings and therefore, must be at the same temperature as its surroundings or pay an entropy cost. These two insights are an important supplement to Smoluchowski's account. But they are no more than a supplement – in many cases Smoluchowski-style exorcisms which make no reference to measurements are still the right ones. One cannot – yet – speak of a new stage in the history of exorcism.

It is also important to notice that not every demon can be described in terms of measurements. In the case of the Smoluchowski trapdoor or the Feynman ratchet and pawl, speaking of measurements would be contrived – the trapdoor can hardly be said to 'measure' the atoms in order to obtain the 'information' whether it has to open or not. And in the case of Zhang and Zhang's demon (Zhang & Zhang [33], 1992), which is simply a non-Hamiltonian force-field, terms like 'information' and 'measurement' are certainly not applicable. Their demon works because it is non-Hamiltonian, and not because it is sensitive to the microscopic state of the gas it works on. Reference to measurements is only useful when the demon under consideration is one which tries to create a correlation between its own microscopic behaviour and that of the gas.

6.2 Information as negentropy

We have seen in the last section that neither Szilard nor Brillouin established a deep connection between measurement and entropy. Nevertheless, their work gave rise to a conception of entropy as somehow opposed to information. The intuition underlying this conception is the following. If Maxwell's Demon only

knew the positions of all the molecules, it would not have to do any measurements. Hence, it would not be exorcised by Brillouin's analysis, and could do its sorting work. So information can be converted into a decrease of entropy. On the other hand, gaining information about the molecules increases the total entropy. These relations are very suggestive, and it is tempting to postulate a conservation law like

$$\Delta \text{ Entropy - } \Delta \text{ Information} = 0,$$

or, more in the spirit of the Second Law,

$$\Delta \text{ Entropy - } \Delta \text{ Information} \geq 0.$$

Another incentive was the mathematical form which C.E. Shannon had given to his measure of amounts of information in his pioneering work on information theory. I will give a brief exposition of his ideas first, before looking at the early application of information theory in connection with Maxwell's Demon. Along the way, it will become clear that the relation between entropy and information which is supposed to be pointed out by the negentropy-analysis is mostly a fiction. Incidentally, information theory will stay with us all the way through chapter 7, where a more recent application of information theory to the demon takes centre stage.

6.2.1 Information and Shannon entropy

In his seminal article (1948, [26]), Claude Shannon set forth a definition of information-theoretic entropy. Consider a set $S = \{S_1, S_2, \ldots, S_n\}$ of n possible events, which are mutually exclusive and exhaustive. The chance of event S_i happening is p_i, with the obvious constraints

$$\sum_{i=1}^{n} p_i = 1 \tag{6.29}$$

and

$$\forall i : p_i \geq 0. \tag{6.30}$$

Can we find a measure of our uncertainty about the outcome? Obviously, if one of the p_i is 1, we have complete certainty. On the other hand, if all of the p_i are equal, the uncertainty is maximal. We wish our measure of uncertainty to have these same properties. Using these and similar constraints,[3] Shannon arrives at the following form for the measure:

$$H(p_1, p_2, \ldots, p_n) = -K \sum_{i=1}^{n} p_i \log p_i. \tag{6.31}$$

Here K is an arbitrary constant, and we are using the logarithm base-2. It is easily checked that H has all the intuitive features of uncertainty. H is also additive: if two independent sets of events, S_1 and S_2, have separate uncertainties H_1 and H_2, the set of joint events $S_{1\&2}$ has uncertainty $H_{1\&2} = H_1 + H_2$.

At this point in his article, Shannon points out an interesting analogy:

[3] See his article for details.

> The form of H will be recognized as that of entropy as defined in certain formulations of statistical mechanics where p_i is the probability of a system being in cell i of its phase space. [...] We shall call $H = -\sum p_i \log p_i$ the entropy of the set of probabilities $p_1, \ldots, p_n$.[4]

Indeed, this is the case. Both Gibbs entropies have the general form $S = -k \sum x \log x$, with an integral replacing the sum for the fine-grained Gibbs entropy. Since the Shannon-entropy, which is the name Shannon gave to H, of a continuous distribution $\rho(\vec{x})$ is defined as

$$H = -\int_{-\infty}^{\infty} \rho(\vec{x}) \log \rho(\vec{x}) d\vec{x}, \tag{6.32}$$

the similarity between Shannon-entropy and statistical mechanical entropy exists for both continuous and discrete distributions. The only difference is that Shannon sets the constants K to 1, whereas in statistical mechanics it is Boltzmann's constant k.

Now suppose that we get some more information about S; perhaps we are told the outcome, or we are told that the outcome lies in a restricted subset of S. Whatever the information may be, if it is non-trivial it will change the values of the p_i. For example, if we are told that outcome j obtains, p_j is set to 1 and all the others to 0. What we would like is a measure of the amount of information which is imparted to us. It is very intuitive to say that the amount of information contained in a message is equal to the decrease of uncertainty. Let P_A be the set of p_i before the information is known, and P_B the set of the p_i after it becomes known. Then

$$\Delta I = H(P_A) - H(P_B) = -\Delta H. \tag{6.33}$$

Which leads to:

$$\Delta(I + H) = 0. \tag{6.34}$$

Given this tantalising equation linking entropy with information, it is hardly a surprise that physicists – Brillouin foremost among them – tried to apply it to Maxwell's Demon. They supposed this lithe being to be defeated by measurements; in other words, by gaining information. Would a careful consideration of the role of information in statistical mechanics perhaps constitute an exorcism of a very general sort?

6.2.2 Information and physical entropy

We now follow Brillouin's discussion (1962, [8]; especially chapters 1, 12 and 13.) of the 'negentropy principle of information'.[5] His treatment differs conceptually from that given in the last subsection, so please read on carefully. Brillouin's basic idea is that initially, we have a system which can be in P_0 states; then, we gain some information about the system, after which it can only be in P_1 states. Assume, says Brillouin, that all states are equally probable, and that in

[4] [26], p. 11. In the line replaced by dots in this quote, Shannon claims that H is, for instance, the Boltzmann entropy. It is evidently not: the Boltzmann entropy is defined for one single state, not for a class of states with associated probabilities.

[5] I ignore his distinction between free and bound information, which is both quite unclear and not very relevant for my purpose.

the initial situation we have no information about the system: $I_0 = 0$. Then the final information I_1 is given by:

$$I_1 = K \log(\frac{P_0}{P_1}). \tag{6.35}$$

Let's motivate this equation with an example. Suppose I have a string of n bits, all of which can be 0 or 1 with equal probability. I do not yet know anything about the string, so I have no information: $I_0 = 0$. The number of possibilities $P_0 = 2^n$. Setting K to 1, formula 6.35 becomes an accurate measure of the number of bits of information I have when I discover more facts about the string. For let me read the first $m < n$ bits; this reduces the number of possible states to $P_1 = 2^{n-m}$. So my information in bits is:

$$I_1 = \log(\frac{2^n}{2^{n-m}}) = \log(2^m) = m, \tag{6.36}$$

which is exactly right. Working out further examples will convince the reader that the above definition of information agrees with our intuitions. By the way, notice what it measures: it measures the information which I have about the system. When I have complete information, I know everything there is to know about the system – namely, in which one state it is.

If we wish to apply these results to thermodynamics, we must set $K = k$. Thus, if $I_0 = 0$

$$I_1 = k \log(\frac{P_0}{P_1}), \tag{6.37}$$

and in general[6]

$$I = I_1 - I_0 = k \log(\frac{P_0}{P_1}). \tag{6.38}$$

The entropy is defined thus:

$$S_i = k \log P_i. \tag{6.39}$$

From formulae 6.38 and 6.39 the following relation between the information I have obtained about a system and the system's entropy is readily deduced:

$$S_1 = S_0 - I, \tag{6.40}$$

and also[7]

$$I = -\Delta S. \tag{6.41}$$

Brillouin's reasoning up to this point is easy to follow, except for one thing: what does he mean with 'entropy'? It appears to have the form of the Boltzmann entropy, formula 2.1, for a discrete state-space where the measure of a macrostate is the number of microstates it contains. But notice that Brillouin is not worrying about macrostates very much: he does not claim that the set of all 'possible outcomes' has to be a macrostate. Nor can he, because presumably it does not have to be – the information one has about a system is not necessarily limited to macrostates. But the sets of possible outcomes, unlike the macrostates, do not form a partition of the state space. The S defined by Brillouin is therefore not equivalent to any measure of physical entropy known

[6] We are still assuming equiprobability of all possible states.

[7] This formula is not given by Brillouin, but it seems a trivial consequence of the last one.

to man. And yet he calls it entropy, and will equate it with thermodynamic entropy a few steps later in his argument. This is conceptually *very* sloppy.

What does Brillouin himself say about the meaning of 'entropy'? Apparently, it is something which changes whenever information about the system is obtained. According to Brillouin ([8], p. 160):

> Entropy is usually described as measuring the amount of disorder in a physical system. A more precise statement is that *entropy measures the lack of information* about the actual structure of the system. This lack of information introduces the possibility of a great variety of microscopically distinct structures, which we are, in practice, unable to distinguish from one another. Since any one of these different microstructures can actually be realized at any given time, the lack of information corresponds to actual disorder in the hidden degrees of freedom.

The problem is that entropy has now become an *epistemic* notion: it measures ignorance. But we are interested in raising weights by cooling objects; we wonder why a ship cannot simply use the heat energy of the sea to get moving; we would like to know why we cannot exploit Brownian motion to do macroscopic work. If entropy is an epistemic notion, it is unclear how it can have anything to do with the occurrence or non-occurrence of such phenomena. One can suggest that perhaps we could create anti-entropic phenomena if we knew precisely the microscopic state of a system. That is a valuable suggestion which will be discussed in subsection 6.3.2. But what Brillouin claims, and what is *essential* for his linking of entropy and information, is *not* that information can be used to lower entropy, but that entropy simply *is* a measure of lack of information as defined by his formula 6.39. Interested as we are in Maxwell's Demon as the creator of objective phenomena, it is hard to see how Brillouin's ideas about 'entropy', information and their interplay is going to tell us anything useful. But let's return to his discussion.

6.2.3 Enter negentropy

Brillouin now remarks that, leaving the system alone after we have gained some information, Carnot's principle states that

$$\Delta S_1 \geq 0. \tag{6.42}$$

Using formula 6.40, we obtain

$$\Delta(S_0 - I) \geq 0. \tag{6.43}$$

This must be one of the most bizarre conclusions ever reached in the physical literature. First of all, although Carnot's principle (the thermodynamical second law) tells us that the *thermodynamical* entropy cannot decrease, it says no such thing about the measure of ignorance Brillouin is pleased to call 'entropy'. In fact, given the fact that our world is Hamiltonian, one would expect that entropy as defined by Brillouin remains constant as long as we don't make measurements – volume in state space is, after all, conserved. S_1, S_0 and I are not functions of time. S_0 measures the entropy of the system just before the observer gains

information I, and S_1 is the value of the system's entropy just after that. It is as meaningless to speak of ΔS_1, ΔS_0 or ΔI as it is to speak of $\Delta 3$.

We now have two incompatible but simultaneously used conceptions of entropy, and three time-independent constants evolving through time. We are ready for anything. At this point Brillouin introduces N, the negentropy of the system, defined by the rather simple relationship

$$N = -S. \tag{6.44}$$

Together with formula 6.43 this yields

$$\Delta(N_0 + I) \leq 0. \tag{6.45}$$

I invite anyone who thinks he still understands what is going on to compare this result with formula 6.41. Whether it is incomprehensible or not, according to Brillouin this discussion and several physical examples which can be found in his book make clear that 'information can be changed into negentropy, and [...] can be obtained only at the expense of the negentropy of some physical system'.[8] This relation can be written as:

$$I \rightleftarrows N, \tag{6.46}$$

although the quantity $I + N$ is not conserved, but can decrease in time as shown by formula 6.45.

This last formula establishes the conceptual connection with Brillouin's torch. The non-equilibrium situation of the hot filament and the cold gas is a state of non-maximal entropy. Left to itself, the filament will radiate its excess energy into the gas, until thermal equilibrium has been achieved. As long as equilibrium has not been achieved, the total system has an amount of negentropy which can be converted into information. *Because* the system is not in its lowest negentropy (highest entropy) state, the demon can get information about the system – but only at the cost of lowering the negentropy, or in other words, raising the entropy. The somewhat plausible idea that information can only be obtained in a non-equilibrium situation, in which case the entropy increase will always keep up with the amount of information gained, is half of the physical intuition underlying Brillouin's ideas about negentropy. It will be discussed in subsection 6.3.1. The other half is the idea that, given enough information about the system, we can lower its entropy: information can be used to break the second law. This will be discussed in subsection 6.3.2.

But Brillouin wants to do more than point out these ideas; he wants to supply us with a rigourous proof of formula 6.45. He wants to show that the acquisition of information *always* implies a corresponding entropy increase; not just in the case of his carefully analysed torch, but in every case conceivable. I have already leveled some criticisms against this proof: it equates two different definitions of entropy, one of which is very hard if not outright impossible to understand physically, and its use of quantities like ΔS_0 defies the imagination. But let me point out some additional problems.

[8] [8], p. 154.

6.2.4 Critique of Brillouin's proof

What elements enter his proof as premisses? Only three: a definition of information in terms of the number of possible cases; a definition of entropy in terms of the number of possible cases (whatever they may be); and Carnot's principle – also know as the second law of thermodynamics.

But wait – the second law of thermodynamics enters as a premiss? Indeed it does, which disqualifies Brillouin's proof as an exorcism of Maxwell's Demon. It is hardly surprising that one can prove that measuring and making use of information cannot lower the entropy of the world, if one uses as premiss that one cannot lower the entropy of the world *at all*. As Earman and Norton (1999 [11], p. 9) point out, Brillouin's proof exorcism succeeds because he assumes what he wishes to prove - but that means that it fails.

There is another task which might be performed by the proof, even though it fails as an exorcism: providing insight in an interesting connection between entropy and information. But it does not carry this task to a successful completion. It is true, of course, that the mathematical equalities between I and S which he writes down at the beginning of his analysis hold. But his S is not a measure of the entropy; in fact, it is totally unclear what the 'number of possible states' of a system is, and thus what S is supposed to mean. If 'the number of possible states *given some information*' is meant, the connection between information and S is clear – but S becomes an epistemic notion rather than an indicator of interesting phenomena. If, using a line of thought going back to Boltzmann, we speak of 'the number of possible microstates given the macrostate', there is a clear connection with the Boltzmann entropy; unfortunately, it is hard to see what this has to do with information. Yes, all the information we can get about a system with macroscopic measurements is in which macrostate it is; but we certainly do not wish to limit the demon to macroscopic measurements! I admit that I cannot imagine a way to interpret Brillouin such that I is an acceptable measure of information *and* S one of physical entropy. I do not think Brillouin has provided us with a compelling reason to believe information and entropy to be closely linked.

The possibility that a better exposition of Brillouin's basic ideas *would* give us this compelling reason remains. There is, however, one very deep reason to be worried about Brillouin's analysis and despair of the possibility of recasting it in a better form. It is the fact that physical reality hardly enters into it. Is it not strange to believe that a link between information-gathering and information-using processes on the one hand, and between temperature, heat and work on the other, can be shown to exist without considering which kind of processes can take place in our world and how they are connected with temperature and other thermodynamical quantities? Suppose that the original demon were not confronted with any background radiation, and could use a very reliable detection-mechanism which involved arbitrarily low amounts of energy exchange. Surely it could gain a lot of information without raising the entropy of the gas or itself – notice that it can stay at an arbitrarily low temperature, since it does not exchange energy with its surroundings. But when Brillouin links entropy and information, background radiation and detection mechanisms play no role whatsoever. In fact, *no* physical considerations come into it. My suggestion is that he is playing an almost trivial game of definition, and forgets to look at the physics which is *really* relevant. Such a strategy cannot possibly

be a success.

6.3 Lessons of measurement and information

As a whole, Brillouin's attempt to formulate a generalised Carnot principle in terms of information and negentropy must be judged a failure. His discussion does not, I believe, establish a deep link between information theory and thermodynamics. Nevertheless, it does draw attention to two interesting ideas concerning information, which I wish to distill from his remarks – as the gold seeker labouriously sifts particles of gold from the mud.

6.3.1 Entropy of measurement

Brillouin's analysis of Maxwell's Demon showed that a torch which emitted black-body radiation is a tool which no self-respecting demon can use. But why would such a statement be generalisable to *all* measurements? Underlying it is perhaps the following idea: "Information can only be obtained in a non-equilibrium situation; but in a non-equilibrium situation, there must be an ongoing entropy increase. This increase is always bigger than the entropy decrease that can be created with the information gained." If we use the terms of Brillouin: "Information can only be obtained by lowering the negentropy of a certain system – for instance, the negentropy readily seen to exist in the case of a cold gas and a hot filament. This negentropy cost defeats the demon, as it is always higher than the entropy reduction he can create with the information." Is it true that acquiring information is only possible by using up some amount of negentropy? Let's split this question into two: is it possible to do a measurement without using an amount of negentropy? And, is it possible to make the measurement apparatus that does this a reliable device? I think the answer to the first question must be a firm 'yes', whereas the answer to the second is less clear but probably negative.

Consider the demon I described which uses a source of radiation at the same temperature as the gas, but with a characteristic peak in intensity at a low frequency. It has been shown that there is simply no entropy increase associated with the detection of an atom using this device. The problem is rather, as I pointed out in subsection 6.1.4, that the detection device using these photons would be subject to thermal fluctuations which defeat it. So it is possible to do a measurement; it is just not possible to do it reliably. This may seem an academic point, but it clarifies the fact that the entropy cost of measurement is not so much an intrinsic property of measuring, as it is a result of the need to differentiate true information from thermal fluctuations.

Let me spell this out again, for it is an important point. Suppose the detecting device has two energy levels, $E = 0$ and $E = \epsilon$. It is at a temperature T, in thermal equilibrium with the surroundings,[9] so the respective probabilities of the lower and the higher energy level stand in the proportion $1 : e^{-\epsilon/kT}$. If the detecting device is to be reliable at all, the energy of the higher level

[9]Remember that it is a fundamental assumption of the entire approach I am discussing that the demon is a system describable with the same kind of temperature as its surroundings. Combined with the inefficient cooling assumption, this implies that it will – after some time – come to be in thermal equilibrium with those surroundings.

must be given by $\epsilon \gg kT$. We also know that two systems in thermal equilibrium exchange, through fluctuations, amounts of energy of order kT. But the detector receives *larger* amounts of energy from the object which supplies the 'negentropy' to do the measurement (the filament, for instance): so there is a heat transfer to the detector. Thus, the demon heats up and endangers its own operation.

This, I think, captures the truth contained in Brillouin's idea of the entropy cost of measurement as expressed in his formula $N \rightleftarrows I$: if we wish to do reliable measurements in order to obtain reliable information, the detector must be heating up. It is essentially the insight arrived at in section 6.1, cast into the form of a potentially misleading terminology. Brillouin does not succeed to prove the second law without assuming the fundamental assumption of subsection 5.3.1, and therefore, he does not succeed in invalidating the conclusions of Part I.

6.3.2 Information and entropy

If Maxwell's Demon had a lot of information about the system on which it has to work, it could work without a problem. This has already been pointed out in section 2.4: if a shutter is programmed to open and close at specific times, there are always some possible initial conditions of the gas for which that sequence is exactly right. If the gas begins in that initial condition, the demon will work. Conversely, for every initial condition of the gas, there is a program which will create the desired reduction of entropy.[10] In a sense, then, having information about the system (knowing its initial condition) enables us to program the shutter in such a way that it operates as intended.

There is a distinction to be made between having the information 'in advance' or gaining it 'through measurement'. A pre-programmed demon can be said to have the information in advance, but it only works on a very specific set of initial conditions. Such demons are not fit to be called 'Maxwell's Demon', we judged earlier. We are therefore interested in demons which, so to speak, start with an open mind about the initial condition of the gas, and gain information about its true state by doing measurements. If a demon can obtain information about the system's state in this way, it can program the shutter accordingly and lower the gas's entropy. In this way there *is* a clear link between information and entropy. But to obtain reliable information, the demon must do accurate measurements, which involve an entropy cost as discussed in the last subsection. It is harder to show that this costs outweighs the benefits in all possible situations, but the present chapter contains arguments which make this result plausible in the case of the temperature-demon. Other demons may follow suit when investigated carefully.

So in the end the real, physical connection between entropy and information comes down to the ability to do accurate measurements. The information-account of the demon's failure gives us, as far as I can see, no insights which were not already contained in the measurement-account; which in turn was only an extension of the fluctuation-account. Specifically, the mathematical

[10] This is not quite true. For *some* initial conditions, almost no gas molecules may come near to the hole at all, no matter how long one waits. This set of initial conditions is assumed to be of negligible size.

equations of Brillouin and others are highly misleading, as they equate very different conceptions of entropy without an adequate justification.

Chapter 7

Erasure: the new paradigm

In the preceding chapter I argued that Brillouin's attempt to pinpoint a connection between measurement, information and entropy was not a success. My diagnosis is that his concrete analysis of measurements using a hot torch is important because it reminds us of the energy exchange that has to take place between any measuring device and its surroundings. That some exchange has to take place is clear; that it has to be so big that it defeats the demon only follows when we make the assumption that every part of the demon and its surroundings is a canonical thermodynamical system. In connection with this, I have suggested that the resolutions of the Szilard-engine and related machines are best sought within the fluctuations-approach. Because of the sensitivity which the measuring instrument must necessarily possess, it either fluctuates wildly and uncontrollably, or it must use entropy-creating sources of energy like Brillouin's filament. In my opinion, there is neither need nor justification for exorcisms which link information processing and entropy in any way but that which I have just described.

This is a minority position. Over the last decades a paradigm has emerged in demonic studies wherein information-processing takes centre stage in exorcisms. This new approach was initiated by Rolf Landauer and staunchly defended by Charles Bennett. Its central claim is that information can be gained without dissipation, but that the demon is defeated in a further stage of its operation: when it resets its memory of the measurement in order to complete the cycle it has to go through. Information erasure must always involve increase of entropy; and this increase always compensates any entropic benefits that one may expect to gain from the measurement done.

In this chapter, we will first look at Landauer's thesis that information erasure leads to entropy increase, and the way in which his principle has been used to exorcise demons. Then we will look into two kinds of proofs of Landauer's principle. The first are explicitly thermodynamical, and are found to be based on the fundamental assumption of subsection 5.3.1. The second try to define a link between information and entropy, an attempt which fails. At best, the proofs of the second class reduce to the State Space Contraction argument of chapter 3. We then discuss an argument by John Norton which tries to show that all attempts to prove Landauer's principle using anything like *SSC* make a fundamental mistake. The final conclusion will be that the erasure-account of exorcism is not particularly enlightening.

7.1 Information erasure

7.1.1 Landauer's principle

Landauer (1961, [17]) first claimed that information erasure, the resetting of a physical memory to a chosen initial situation, is necessarily accompanied by an increase of entropy. A computer-scientist working for IBM, Landauer was interested in dissipation in computing devices. It had been suggested that every step in a computation generated heat, thus increasing the temperature and the entropy of the computer. According to Landauer, however, every computational step can – in principle – be done in a dissipationless way, except for those that are logically irreversible.

A computational step is reversible if its input can be reconstructed from its output. The negation is reversible: if the result of a negation is 1, the input was 0; if it is 0, the input was 1. An `AND`-operation is not reversible: its output is merely one bit, from which its two input bits can never be reconstructed. It is possible to do every computation in a logically reversible way if a sufficient amount of memory is available. In the case of the `AND`-operation, which takes two bits as input, one would need three bits of output: the first containing the result of the `AND`-operation, the other two containing for instance the original bits. From these it would be trivial to recreate the input: hence, the augmented `AND`-operation is logically reversible.

At the end of such a computation, if the computer is supposed to work in a cycle, it should erase its memory. It can easily be seen that this is a logically irreversible step. After erasure, the memory is in one predesignated state, perhaps all zeroes. The erasure procedure maps every state in which the memory could possibly be onto this ground state, so the state before erasure cannot be calculated from the state after erasure. Erasure is logically irreversible.

According to Landauer, there is a connection between logical irreversibility and entropy increase. In particular, he asserts

Landauer's principle: The erasure of one bit of information is accompanied by a minimum average entropy increase of $kT \ln 2$.

To see the initial plausibility of the idea that logically irreversible operations must generate entropy, imagine a particle in a double well potential, where the two minima are labeled `ZERO` and `ONE`. Between the minima is a potential barrier of height akT, with $a \gg 1$ to ensure that the memory is reliable. The temperature T is that of the bit's environment – the computer, say – which is thought of as an infinite heat bath. Suppose we wish to apply the operation 'Restore to `ZERO`', which takes the particle to `ZERO` no matter where it is now. There are two possibilities: either we apply a different force on the particle depending on its current position, or we apply the same force to the particle independently of its current position.

In the first case, dissipation is not necessary. If the particle is at `ZERO` already, we simply apply no force at all. If it is at `ONE`, we can apply a force to push it over the boundary which separates the two minima, then apply a retarding force so the particle will not have any kinetic energy left when it reaches `ZERO`. During the particle's descent we can regain all energy we put into it during its ascent, and there has been no increase of entropy.

But this first case is not an example of true memory erasure. A computer which resets its memory must do this in a manner independent of the exact data being handled. For suppose the erasure procedure to be sensitive of the data it works on. This means that the device that carries out the erasure must take one of two possible courses of action, depending on the bit it is erasing. But then the erasing device is itself a memory, the path it takes exactly representing the original bit of data. This means that at the end of the resetting operation, the erasing device itself contains the information we wished to erase, and must itself be reset – perhaps by some second device, and we end up with an unending chain of erasing devices.

Can memory erasure be done in a dissipationless manner? Reconsider the way in which we brought a particle from `ONE` to `ZERO`. First we applied a force to get it to the top of the potential barrier. At that point is has a potential energy akT. Then, as it moves downwards towards `ZERO`, the potential energy is changed into kinetic energy, which in turn drops as the particle moves against the force – now applied the other way – so that it comes to rest at the bottom of the well. What happens if we apply this same force to a particle already at `ZERO`? First, it will climb the leftmost infinite potential barrier to a height akT. Then the force towards the left is changed into a force towards the right, and the particle will have gained an amount of kinetic energy higher than akT when it reaches `ZERO`; it will spend the rest of dissipationless eternity oscillating back and forth between `ZERO` and `ONE`. The erasure is clearly not successful.

In a dissipationless environment there is no force which erases the information yet can be applied to the particle regardless of its current position. This is true because in a non-dissipative system governed by Hamiltonian equations of motion, volume in state space is conserved. The set of particle-states corresponding to `ZERO` or to `ONE` cannot be mapped onto the set containing only the states corresponding to `ZERO` in a conservative way. Such a mapping would be a compression of the accessible volume in state space by a factor of two. This is impossible without dissipation.

In the example we discussed, the amount of energy to be dissipated is akT. The amount which has been accepted as the theoretical minimum average dissipated energy for any bit-erasure whatsoever is $kT \ln 2$, with an associated entropy increase of $k \ln 2$. Alleged proofs of this minimum entropy increase will be discussed in sections 7.2 and 7.3. First, we will describe how Landauer's principle can be used to exorcise the demon.

7.1.2 The erasing demon

We return to the Szilard engine described in subsection 6.1.1. Using the one-molecule gas to push a piston and raise a weight, the demon is able to extract an amount of heat $kT_1 \ln 2$ from a heat bath at temperature T_1, transforming all of it to work. This process lowers the entropy of the heat bath by $k \ln 2$, seemingly without raising the entropy of anything else in the world. Szilard claimed that measuring the position of the particle has an intrinsic entropy cost of $k \ln 2$ associated with it, which saves the second law. The advocates of Landauer's principle deny that measurement generates entropy: according to them it is a logically reversible transformation which therefore can, in principle, be carried out without any entropy cost. The demon is defeated in a different way. After it has made a measurement, the memory of the demon contains the result of the

measurement. This is one bit of information: either the particle is on the left or it is on the right. The demon must erase this bit at the end of his cycle. And this is where Landauer's principle comes in: erasing this bit is accompanied by an entropy increase of $k \ln 2$, exactly enough to save the second law from the demon.

One may wonder why the demon cannot simply forgo erasing its memory-bit, and let it be overwritten by the next measurement it makes. The process of overwriting, however, is effectively an erasure. Suppose the demon finds the particle in the left half of the box. Then it has to perform the operation 'Set to `LEFT`', regardless of its current state. This is the equivalent of an erasing procedure, and carries the same entropy cost. There are no two ways about it: the demon must reset its memory, so if Landauer's principle is correct, it constitutes an effective exorcism of the demon. But is it correct? Most of the rest of this chapter will be dedicated to answering this question.

7.1.3 Exorcist XIV: the dilemma

In the second part of their critical assessment of the Maxwell's Demon literature *Exorcist XIV: The Wrath of Maxwell's Demon* (Earman & Norton 1999, [11]), John Norton en John Earman argued vehemently against the information-theoretic exorcisms using Landauer's principle.

> Our thesis in this paper is that information theoretic analyses provide largely illusory benefits: they are either essentially trivial restatements of earlier presuppositions or posits without proper foundation.[1]

To 'sharpen' the thesis, they formulated a dilemma. Any information-theoretic exorcism must choose between two horns of the dilemma: the 'sound' horn and the 'profound' horn. Consider the combined system of the Demon and the system on which it is to operate. Either this combined system is a canonical thermal system, or it is not. If it is, the second law obtains for the combined system, so the Demon cannot succeed and is defeated trivially. No discussion of information erasure is necessary, and the exorcism is sound, but trivial rather than profound. If, on the other hand, the combined system is not a canonical thermal system, then the information theoretic exorcisms might be able to become profound: they would have to prove a new and independent principle strong enough to exorcise Maxwell's Demon. Unfortunately – so say Earman and Norton – no such proof can be found in the large quantity of material written about Landauer's principle and related subjects. Thus, all discussion to date fail because they are either unsound (not proven), or not profound (trivial extensions of earlier assumptions).

Before we can assess the plausibility of this dilemma actually obtaining, we'll have to look at a number of proposed proofs for Landauer's principle. These can be divided into two classes: those based on thermodynamics alone, and those which invoke principles of information or volume in state space which are not found in orthodox thermodynamics. It is my contention that these two classes correspond, at least in a rough an ready way, to Earman and Norton's sound

[1] [11], p. 2.

and profound horns of the dilemma. We will look at the first kind of proof in section 7.2, at the second kind in section 7.3.

7.2 Proofs of the first kind

7.2.1 Landauer's proof

The first kind of proof concerns itself with thermodynamical models only, without formulating and trying to prove deeper lying, completely universal principles connecting information and the Demon. Nevertheless, according to Leff and Rex (2003, [20], p. 34.) they do provide the answer to the dilemma of Earman and Norton. Earman en Norton were "evidently unaware of the proof of Landauer's principle by Shizume (1995)", and their work "also preceded the proofs (classical and quantum mechanical) by Piechocinska (2000)". We will shortly investigate these alleged proofs, and wonder whether they are sound, and whether they are profound. But let us start by taking a look at the original article by Landauer. Rolf Landauer's 1961 paper ([17]) contains a discussion of a particle in a bistable potential well. Memory erasure is effected by creating a potential difference between the two minima and letting the Brownian motion of the particle take it to the side with the lowest potential. The argument for Landauer's principle – not yet so called, of course – is surprisingly short and simple:

> Consider a statistical ensemble of bits in thermal equilibrium. If these are all reset to `ONE`, the number of states covered in the ensemble has been cut in half. The entropy therefore has been reduced by $k \log_e 2 = 0.6931k$ per bit. The entropy of a closed system, e.g., a computer with its own batteries, cannot decrease; hence this entropy must appear elsewhere as a heating effect, supplying $0.6931kT$ per restored bit to the surroundings. This is, of course, a minimum heating effect, and our method of reasoning gives us no guarantee that this minimum is in fact achievable.[2]

Immediately evident is that Landauer *assumes* the validity of the second law of thermodynamics: 'the entropy of a closed system cannot decrease'. If used as an information-theoretic exorcism it evidently chooses the 'sound' horn of the dilemma, renouncing all claims to profoundness. Unless Landauer's principle is supported by arguments which do not assume what an exorcist wishes to prove – that the second law is safe from demons – it cannot be used in a profound, interesting, or even non-trivial way to exorcise the demon. The task that Shizume and Piechocinska took upon themselves was to furnish such arguments.

7.2.2 A thermodynamical proof

Shizume 1995 ([27]) aims to show that for "a system including a particle making the Brownian motion in a time-dependent potential well" Landauer's principle holds rigourously "if the random force acting on the particle is white and Gaussian". Piechocinska 2000 ([24]) derives Landauer's principle microscopically "for a classical system with continuous space and time, with discrete space and time,

[2] Landauer 1961, [17]; reproduced in Leff and Rex 2003, [20], p. 152.

and for a quantum system". Both articles are quite dense, and it seems of little use to recapture all arguments here, especially since the details are not of great importance to us. Therefore, we will only look at Barbara Piechocinska's first proof, for a classical system with continuous space and time.

We have a double well potential, symmetric around $x = 0$ described by the function $U(x)$. The state `ONE` corresponds with $x > 0$, and the state `ZERO` with $x < 0$. We are going to look at an ensemble of bits, and assume that they are in contact with a thermal reservoir of temperature T. Then, the initial ensemble can be described by:

$$\rho_I(x,p) = \frac{1}{Z}\exp\{-\beta[U(x) + \frac{p^2}{2m}]\}, \tag{7.1}$$

where $\beta = 1/(k_B T)$. The ensemble after erasure will be:

$$\rho_F(x,p) = \begin{cases} \frac{2}{Z}\exp(-\beta[U(x) + \frac{p^2}{2m}]) & \text{for } x > 0 \\ 0 & \text{for } x < 0. \end{cases} \tag{7.2}$$

The total system – bit and heat reservoir – is isolated and classical, so it evolves according to the Hamiltonian

$$H(x,p,\vec{x}_T,\vec{p}_T,t) = H(x,p) + H_T(\vec{x}_T,\vec{p}_T) + H_{int}(x,p,\vec{x}_T,\vec{p}_T), \tag{7.3}$$

where the time t is suppressed in every term on the right side of the equation, $H(x,p)$ denotes the Hamiltonian of the bit, H_T that of the heat reservoir, and H_{int} is the interaction term. We define $\zeta = (x,p,\vec{x}_T,\vec{p}_T)$. Then $\zeta(t)$ is a trajectory that described the evolution of all degrees of freedom of a combined system. Assume the erasure process takes a time τ, and define $\zeta^0 = \zeta(0)$ and $\zeta^\tau = \zeta(\tau)$. We now define the function Γ:

$$\Gamma(\zeta^0,\zeta^\tau) = -\ln[\rho_F(x^\tau,p^\tau)] + \ln[\rho_I(x^0,p^0)] + \beta\Delta E(\vec{x}_T^0,\vec{p}_T^0,\vec{x}_T^\tau,\vec{p}_T^\tau), \tag{7.4}$$

where $\Delta E(\vec{x}_T^0,\vec{p}_T^0,\vec{x}_T^\tau,\vec{p}_T^\tau) = H_T(\vec{x}_T^\tau,\vec{p}_T^\tau) - H_T(\vec{x}_T^0,\vec{p}_T^0)$ is the change in the internal energy of the heat reservoir. Γ has no easy intuitive meaning, it is just so defined for computational purposes. Because the evolution is deterministic, ζ^τ is actually a function of ζ^0, and Γ can therefore be seen as a function of ζ^0 alone.

In order to prove Landauer's principle, we must find an inequality about the average heat released into the environment in a transition from the ensemble ρ_I to ρ_F. To get there, we will compute $\langle\exp(-\Gamma)\rangle$, where the angular brackets denote the ensemble average.

$$\begin{aligned} \langle\exp(-\Gamma)\rangle &= \frac{1}{Z_T}\int \rho_I(x^0,p^0)\exp\left(-\frac{H_T(\vec{x}_T^0,\vec{p}_T^0)}{k_B T}\right) \\ &\quad \times \exp(-\Gamma)d\zeta^0 \qquad (7.5) \\ &= \frac{1}{Z_T}\int \rho_I(x^0,p^0)\frac{\rho_F(x^\tau,p^\tau)}{\rho_I(x^0,p^0)} \\ &\quad \times \exp\left(-\frac{H_T(\vec{x}_T^0,\vec{p}_T^0)}{k_B T}\right) \\ &\quad \times \exp\left(\frac{H_T(\vec{x}_T^0,\vec{p}_T^0)}{k_B T} - \frac{H_T(\vec{x}_T^\tau,\vec{p}_T^\tau)}{k_B T}\right) d\zeta^0 \qquad (7.6) \end{aligned}$$

$$= \frac{1}{Z_T} \int \rho_F(x^\tau, p^\tau) \exp\left(-\frac{H_T(\vec{x}_T^\tau, \vec{p}_T^\tau)}{k_B T}\right) d\zeta^\tau \tag{7.7}$$

$$= \frac{Z_T}{Z_T} = 1. \tag{7.8}$$

We changed the integration variable from $d\zeta^0$ to $d\zeta^\tau$, but because the evolution is Hamiltonian the associated Jacobian is 1. We now have $\langle \exp(-\Gamma) \rangle = 1$, which implies, because of the convexity of the exponential function:

$$-\langle \Gamma \rangle \leq 0. \tag{7.9}$$

Explicitly, this becomes:

$$\langle \ln[\rho_F(x^\tau, p^\tau)] \rangle - \langle \ln[\rho_I(x^0, p^0)] \rangle \leq \langle \beta \Delta E \rangle. \tag{7.10}$$

If we put the distribution functions of formulae 7.1 and 7.2 into this equation, and take several mathematical steps which I shall spare the reader, we arrive at

$$\ln(2) \leq \beta \langle \Delta E \rangle. \tag{7.11}$$

We assume that the interaction energy is negligible, so that we can write the law of conservation of energy thus:

$$W = \Delta E + \Delta E_B, \tag{7.12}$$

where W is the work done in the combined system, ΔE is the change in internal energy of the heat reservoir, and ΔE_B is the change in internal energy of the bit. Because $U(x)$ is symmetric, erasure will not change the average energy of the bit: $\langle \Delta E_B \rangle = 0$. So, equations 7.11 and 7.12 together give us:

$$k_B T \ln(2) \leq \langle W \rangle. \tag{7.13}$$

The work performed on the combined system, which is the total heat added to the heat reservoir, must on average be at least equal to $k_B T \ln(2)$ per bit of information. We have derived Landauer's principle.

7.2.3 Critical discussion

Deriving Landauer's principle is quite a feat. Two questions of the highest importance must be answered about the proof presented in the previous section: Is it sound? And what are the presuppositions which are needed to derive it? We will tackle the second question first. The great worth of Piechocinska's article lies not only in her elegant proof of Landauer's principle, but also in her explicit statement of the presuppositions needed to derive it. She identifies the following assumptions:

1. Our system is classical.
2. The memory state is a symmetric double potential well where the states "zero" and "one" have the same energy before and after the erasure.
3. The input is randomly distributed (the number of "zeros" and "ones" is equal and there are no correlations between the bits).
4. During erasure the system is in contact with a thermal reservoir with initial states chosen from a canonical distribution.
5. The interaction term in the Hamiltonian is negligibly small.[3]

[3] [24], p. 062314-6.

Of these, the first is uninteresting, since it delimits the domain of physics to which we restrict ourselves. The second is a description of the memory-device under consideration; it threatens the generality of the proof. But I know of no reasons why the equal-energy bistable potential well would be particularly misleading. So while it is good to keep in mind that Piechocinska's proof is less than completely general, if successful it is nevertheless remarkable. The third assumption actually says that the 'bit' indeed contains one bit of information – unequal numbers of zeroes or ones, or correlations between bits, would diminish the information content of the system to less than a bit per double potential well. Since Landauer's principle states that the erasure of one bit of information must be accompanied by an entropy increase of $k \ln 2$, assuming that we are actually talking about one bit of information is making not a restrictive, but a necessary, assumption. The fourth is the most significant assumption, to which we will return shortly. As to the fifth, not neglecting the interaction energy will change formula 7.12 to

$$W = \Delta E + \Delta E_B + \Delta E_{int}, \tag{7.14}$$

and therefore formula 7.13 to

$$k_B T \ln(2) \leq \langle W - \Delta E_{int} \rangle. \tag{7.15}$$

If ΔE_{int} is negative – that is, if the interaction energy is higher before than after erasure – Landauer's principle can be broken. But it is actually hard to see why it matters much to Piechocinska's discussion whether ΔE, the change in internal energy of the heat bath, is furnished by the environment doing work, the interaction energy decreasing, or some combination of the two. If ΔE changes by at least $k_B T \ln 2$, this counts as an energy of $k_B T \ln 2$ being dissipated into the environment of the bit. And this conclusion is already reached in formula 7.11. I conclude that assumption 5 is not crucial to Piechocinska's purposes.

That leaves assumption 4 as the most important one. "During erasure the system is in contact with a thermal reservoir with initial states chosen from a canonical distribution," she writes. But actually it is a little stronger: "During erasure the system is in contact with *nothing but* a thermal reservoir obeying the canonical distribution, since the combination of the two follows a Hamiltonian evolution". One may wonder how external work can be done upon the system if it is completely isolated, but we will cover that minor inconsistency with the cloak of charity. It is more important to recognise that the stronger version of assumption 4 is simply a guise for the *extended fundamental assumption* of subsection 5.3.1; and the extended fundamental assumption assumed all the contingent facts I have been pointing out in Part I. Barbara Piechocinska, then, has produced a beautiful proof of Landauer's principle, but only at the cost of making the extended fundamental assumption. Her proof is sound, but only profound within a limited context. It cannot be hailed as a general proof of the non-existence of Maxwell's Demon.

7.3 Proofs of the second kind

It is now time to turn to a second kind of proof, which makes more explicit use of general statements on information and its relation with heat or entropy,

and relies less on pure thermal physics. No single argument for the connection between information erasure and the defeat of Maxwell's Demon has been accepted by all proponents of Landauer's principle. I will therefore present three slightly different positions. First, Bennett 1982 ([2]) bases Landauer's principle on the idea that a compression of the occupied volume in state space of a system must be accompanied by an entropy increase elsewhere. Then, Bennett 2003 ([6]) introduces the distinction between information bearing degrees of freedom and non information bearing degrees of freedom, which he exploits to justify Landauer's principle. But he also, in silent conflict with his earlier remarks, interprets entropy as a measure of 'subjective' knowledge. I will argue that the underlying idea of these attempts at exorcism is the State Space Contraction argument of chapter 3, which was refuted in Part I of this thesis. Embellishing it with the concept of information does not help; adding the notion of knowledge to the mix only serves to dissolve the discussion into incoherence.

7.3.1 State space compression

The simplest argument for Landauer's principle appears in part 5 of Charles Bennett's 1982 article ([2]). He asks us to imagine Szilard's engine, as described in subsection 6.1.1. A single molecule flies about in a container which is always in contact with a heat bath at temperature T. The container has two sides, a left side and a right side, which have the same size. We will call the part of the molecule's state space accessible to it V. $V = V_L + V_R$, where V_L and V_R correspond to the left and the right side of the box, respectively. The demon's memory is in a single state, S, at the beginning of the cycle. It's memory has two other states, L and R, for a total of three possible states. We will use the notation (X, Y) to designate the part of the total state space spanned by X in the molecule's state space and Y in the memory's. Szilard's engine now operates in the following five steps.

1. The molecule wanders freely through the container, while the demon's memory is in state S, indicating that it does not know where the molecule is. The system is in (V, S).

2. A partition is inserted into the container, effectively trapping the molecule in either the left or the right half, with a 50% change of either. The system is still in (V, S), so there is no change of volume in state space.

3. The demon performs a measurement on the particle to find out where it is. If it is on the left side, the demon's memory will be set to state L, if it is on the right side, to state R. The system is now in $(V_L, L) \cup (V_R, R)$, which still has the same volume.

4. The demon uses his knowledge of the molecule's position to put a piston on the other side of the box, and raises the partition. The molecule will now do work against the piston, until is expanded to its original volume again, transforming $kT \ln 2$ Joule of heat from the heat bath into work. At the end of this phase, the molecule can once again be anywhere in the container, but the demon's memory has not yet been erased. The system is in $(V, L) \cup (V, R)$, which is twice as big as the part of state space from which it started out.

5. In order to complete the cycle, the demon must reset its memory, which implies a twofold compression of the occupied volume in state space. This "cannot be made to occur spontaneously except in conjunction with a corresponding entropy increase elsewhere. In other words, all the work obtained by letting the molecule expand [in step 4] must be converted into heat again in order to compress the demon's mind back into its standard state."[4]

Bennett's treatment of the Szilard engine in unproblematic, and his talk about ensembles also makes sense in the light of our discussion of ensembles in subsection 2.4.1. It shows that if a demon operates Szilard's engine, its memory erasure is equivalent to a compression of the ensemble's volume in state space by a factor two. But as a proof that memory erasure must be accompanied by entropy increase, this story is still severely lacking. That compression of volume in state space must be accompanied by an entropy increase is by no means clear. That *all* the work obtained by letting the molecule expand must be reconverted into heat is even less clear. What justification does Bennett offer for his claims?

Actually, he does not offer any justification. The nearest he comes to it is by saying that "logically irreversible operations must be avoided entirely in a ballistic computer, and for a very simple reason: the merging of two trajectories into one cannot be brought about by conservative forces."[5] Barring the part about ballistic computers, this reminds us of the way in which the State Space Contraction argument used the Hamiltonian nature of the world as an argument for demanding that volumes in state space remain conserved. Such reasoning could be applied to Bennett's argument as well. It would prove that if the memory is erased, the volume in state space of some other system must be increased. But this is far weaker than Bennett's claim that memory erasure must be accompanied by a corresponding *entropy increase*. To prove this latter claim, we would need a definition of entropy which is relevant for purposes of judging the demon, and a proof that this measure of entropy always increases if the volume in state space of an ensemble increases. Neither can be found in Bennett's paper.

7.3.2 Information bearing degrees of freedom

Bennett 2003 ([6]) remedies the most apparent weaknesses of his earlier account by creating a connection between information and a version of the State Space Contraction argument. Just as there are macroscopic and microscopic degrees of freedom in statistical physics, he notes, so can one speak about two sorts of degrees of freedom in a computer. There are *information bearing degrees of freedom*, IBDF, which are used to encode the logical state of the computation. By design they are sufficiently robust that the computer's logical state evolves deterministically as a function of its initial value, regardless of small fluctuations in the computer's environment or its other degrees of freedom. These latter are called *non-information bearing degrees of freedom*, NIBDF. By definition every degree of freedom of the computer is either an IBDF or an NIBDF. We will quote Bennett:

[4] [2], p. 307.
[5] p. 300.

> While a computer as a whole (including its power supply and other parts of its environment), may be viewed as a closed system obeying reversible laws of motion (Hamiltonian or, more properly for a quantum system, unitary dynamics), Landauer noted that the logical state often evolves irreversibly, with two or more distinct logical states having a single logical successor. Therefore, because Hamiltonian/unitary dynamics conserves (fine-grained) entropy, the entropy decrease of the IBDF during a logically irreversible operation must be compensated by an equal or greater entropy increase in the NIBDF and environment. This is Landauer's principle.[6]

In a few simple sentences, Bennett has connected information with entropy, and proposed an elegant and simple proof of Landauer's principle. Let us take it apart one bit at a time.

First, the computer's degrees of freedom are distinguished as either information bearing or non-information bearing. For a typical memory device such as we have seen in discussions of Maxwell's Demon, this division is rather clear. It does restrict the applicability of the erasure-approach to exorcism to the class of systems for which the division between IBDF's and NIBDF's can be easily drawn, but this is not necessarily a problem. It merely shows that Landauer's principle cannot be the final answer to questions of exorcism, and cannot be used to banish every type of demon.

Second, the computer is seen as a closed system which is governed by Hamiltonian (or unitary, but this does not really matter) dynamics. It is then argued that fine-grained entropy is conserved by Hamiltonian evolution, and that any decrease of it in the IBDF's must be compensated by a corresponding increase in the NIBDF's. This reminds us of *SSC*, see section 3.2, but with two major differences. These differences will be commented upon in points three and five.

Third, instead of using volume in state space and the fact that this is conserved, the argument uses the fine-grained Gibbs entropy and the fact that it too is conserved. For the validity of the argument, this does not matter, but Bennett wishes prove something about the non-occurrence of what I have called 'anti-entropic phenomena'. The fact that the fine-grained entropy is conserved or distributed among certain degrees of freedom implies in itself *nothing* about phenomena, as has been argued in section 2.3. Using the word 'entropy' may seem to make the discussion relevant for the demon, but until it has been shown how the fine-grained Gibbs entropy in this argument hangs together with phenomena, this relevance is not established.

Fourth, Bennett says that there must be "an equal or greater entropy increase" as compensation. This is technically true but in practice misleading, since the entropy increase is necessarily equal, never greater. It is, after all, conserved. Since we know that the thermodynamical entropy can in fact increase, this shows once again that there is a discrepancy between the fine-grained Gibbs entropy and the thermodynamical entropy we would rather have. A malicious interpreter might think Bennett tries to conceal this by the sly trick of changing equality to 'equal or greater than'. We are not malicious interpreters.

Fifth, instead of using microscopic and macroscopic degrees of freedom like *SSC*, Bennett's argument uses information bearing and non-information bearing

[6] [6], p. 502

degrees of freedom. Unfortunately, where the microscopic/macroscopic distinction could easily be linked to the phenomenal definition of entropy (more energy in the microscopic degrees of freedom means more heat), this is not so clear for the IBDF/NIBDF distinction. Why would removing energy from information bearing degrees of freedom to non-information bearing degrees of freedom imply an increase of heat? Bennett recognises that it does not: "Typically, the entropy increase takes the form of energy imported into the computer, converted to heat, and dissipated into the environment, but it need not be, since entropy can be exported in other ways, for example by randomizing configurational degrees of freedom in the environment".[7] But if this is so, how does Landauer's principle defeat Maxwell's Demon?

Sixth, if the problem that information and fine-grained Gibbs entropy as such seem to have little relevance to Maxwell's Demon is solved by equating IBDF's to macroscopic and NIBDF's to microscopic degrees of freedom – and some such step seems necessary –, we simply have *SSC*. The weaknesses of *SSC* have been shown in chapters 3 and 4. If it is the core of the new erasure-approach to exorcism, as I have stated when I presented it and as seems evident from the present discussion, the erasure-approach cannot be judged a success. It does not banish Maxwell's Demon.

7.3.3 Subjective information

Several lines below his argument for Landauer's principle which was discussed in the previous subsection, Bennett makes a curious comment.

> If a logically reversible operation like erasure is applied to random data, the operation still may be thermodynamically reversible because it represents a reversible transfer of entropy from the data to the environment [...]. But if, as is more usual in computing, the logically irreversible operation is applied to known data, the operation is thermodynamically irreversible, because the environmental entropy increase is not compensated by any decrease of entropy of the data. This wasteful situation, in which an operation that *could* have reduced the data's entropy is applied to data whose entropy is already zero, is [...][8]

Somehow, according to Bennett, the thermodynamical properties of an operation depend on our knowledge. Yet surely, whether a demon can or cannot operate has nothing at all to do with our knowledge about it or the gas on which it is meant to operate? The only way to inject the concept of knowledge into the story of the demon is by interpreting the ensemble of systems, as characterised by the distribution function $\rho(\vec{x})$, as a measure of our ignorance about the system. We do not know in which state a single system is, all we know is that it has a certain chance of being in a certain state – knowledge which is summarised in the function $\rho(\vec{x})$. So if we do not know what data we are going to erase, $\rho(\vec{x})$ will be evenly distributed over all possibilities. Erasing this data will decrease the fine-grained entropy of our memory device. But if we already know what state the memory is in, if we already know the data, $\rho(\vec{x})$ will be

[7] [6], p. 502.
[8] [6], p. 502.

completely concentrated in one point and the fine-grained entropy cannot be lowered by memory erasure. But because the *operation* of erasure is identical in both cases, the entropy it generates must also be the same. So if we erase known data, we generate entropy without getting anything in return.

This reconstruction of Bennett's reasoning at least explains his remarks; but it cannot save them from heavy criticism. Most obviously, his claim that erasing known data increases the total entropy must be rejected: the fine-grained Gibbs entropy cannot increase. There are no two ways about it, it must always remain constant. If there is no multiplicity in the IBDF's, there cannot be an increase in the NIBDF's.

A second criticism concerns the paradoxical consequences that Bennett's view has. If we erase known data the entropy increases, no matter what the actual data is. But the erasure process is not dependent on our knowledge, so from a physical perspective our knowledge must be irrelevant: we just have some data and a process (erasure) acting upon it. This cannot be physically different from erasing unknown data, the only difference being that we do not know what exactly is taking place. But physically, everything is the same: the same processes are taking place, whether the data is known or unknown. So if entropy increase depends on our knowledge of the data, this entropy cannot be a *physical* quantity. And if it is not a physical quantity, it cannot have anything to do with Maxwell's Demon.

Thirdly, it seems pretty hard to argue for the view that our knowledge, formalised as $\rho(\vec{x})$, must follow a Hamiltonian evolution. And finally, there is a perfectly good reason why we must use ensembles in our discussion of Maxwell's Demon. We do not want it to be overly sensitive to initial conditions, an argument which was developed in subsection 2.4.1. But that we do or do not have knowledge about a certain system is *not* a good reason for using ensembles. Our knowledge is nothing to the demon, and certainly not the reason he must act on an ensemble instead of on a single system.

For these reasons, I conclude that interpreting the ensemble $\rho(\vec{x})$ as a measure of knowledge does not help the erasure argument against Maxwell's Demon. In fact, it only makes it very hard to understand the sound physical ideas beneath the misleading statements at the surface. Bennett is not alone in bringing knowledge into the story. See Jeffrey Bub 2000, [9], for another example. It is very amusing that Bub provides an argument (on page 6 of his article), using simple thermodynamical thought experiments with which one can hardly disagree, that it is only erasing *unknown* information that must be accompanied by entropy increase. This is the *exact* opposite of Bennett's conclusion, which suggests that Bennett's approach to information, subjective knowledge and entropy may be fundamentally confused. There does not seem to be much hope for a deep argument for Landauer's principle in this direction. All in all, then, the exorcisms of the second kind are profound, but not sound – although they do not presuppose the fundamental assumption, they do not prove the non-existence of Maxwell's Demon.

7.4 Too many notions of 'ensemble'

In this section, we will look at a further important criticism of the erasure-school of exorcism: John Norton's claim that the arguments for Maxwell's De-

mon which rely on erasure implying a contraction in state space are based on illegitimate assumptions. I will argue that his verdict is too harsh, and that a confusion about which notion of 'ensemble' to use is the root of the controversy.

7.4.1 Eaters of the Lotus

In his 2004 article *Eaters of the Lotus* ([23]), John Norton claims that Bennett's argument of subsection 7.3.2 fails. Remember that Bennett argued that since fine-grained entropy is preserved by Hamiltonian evolution, and the resetting of a memory is a many-to-one mapping which decreases the fine-grained entropy of the memory, the entropy of the external world must be raised during erasure. We identified his argument as a somewhat misleadingly stated version of the State Space Contraction argument, which, although a good argument, can be countered effectively. But according to Norton Bennett's 'Many to One Mapping Argument', *MOMA* from now on, fails because it makes an illicit use of thermodynamical ensembles. It is not a good argument the presuppositions of which can be successfully contested, it is a bad, confused argument.

Norton starts by telling us about the two different ways in which ensembles can be constructed in statistical mechanics. I will quote him at some length:

> One familiar way of doing this is to take a single component and sample its state frequency through its time development. The probability distribution of the component at one moment is then recovered from the occupation times, the fractional times the system has spent in different parts of its phase space during the history sampled. [...] Another way of doing it is to take a collection of identical components with the same phase space – an "ensemble" – and generate a probability distribution in one phase space from the relative frequency of the positions of the components in their own phase spaces at one moment in time. [...] We might take the probability distributions of one component at different times; or we might take the probability distributions of many components from their phase spaces. Carried out correctly, this form of the procedure is rather trivial, since all the distributions are the same. In all cases, the result is a probability distribution in one phase space at one moment that represents the thermodynamic properties of one component.[9]

It is important to note that Norton claims that the probability distribution as a time-average of a single component and the probability distribution as the description of an ensemble must be exactly equal – *if* the procedure is carried out correctly. This is not so much a presupposition as a restriction on what Norton accepts as legitimate "ensembles": a legitimate ensemble is one in which all components are described individually by the probability distribution. The probability distribution the characterises the ensemble must also be the time-averaged position in state space of every system in the ensemble.

With this idea firmly in place, John Norton goes on to criticise the use of ensembles in exorcisms inspired by Maxwell's Demon. It is a presupposition of arguments such as Bennett's that a memory in which unknown data is stored, which thus has equal chances of being in a `ONE` or a `ZERO` state, is characterised

[9] [23], P. 14-15.

by a distribution function which occupies a *larger* volume in state space than that of a reset cell, which is in `ZERO` with certainty. If the memory is a one-molecule gas in a partitioned box, the distribution function should range over the part state space that corresponds to the `ZERO` side of the box and the part of state space that corresponds to the `ONE` side. But if the memory is reset, the distribution function should only range over the part of state space that corresponds to the `ZERO` side.

According to Norton, this idea is wholly incoherent, because it makes use of an illegal ensemble: an ensemble where some of the systems have a molecule trapped on the `ZERO` side, and some have a molecule trapped on the `ONE` side. If we sample the state frequency of one of the former systems through its time evolution, we will get a distribution function confined to the `ZERO`-part of state space; if we do the same for of the latter, we will get a distribution function confined to the `ONE`-part of state space. And if we take the ensemble average, we get a distribution function evenly spread out over both parts. Thus, the distribution function read off from the ensemble is not equal to the time-average of *any* system within the ensemble, and does not represent "the thermodynamic properties of one component". The ensemble is illicit, because it has nothing to do with the physical reality of the single system or its thermodynamical state.

Adopting these kinds of ensembles even entails giving up the additivity of entropy, claims Norton. For suppose we have a set of N systems in state `ZERO`, and assume that the corresponding distribution function leads to a thermodynamic entropy of S; the total entropy of this set is NS. We also have an equally large set of systems in state `ONE`, which is just as big in state space as state `ZERO`, such that the total entropy of this set is also NS. What happens if we put the two sets together into a new ensemble? The new distribution function will be spread out over a volume in state twice as big as the original one for either set, so the entropy of each component becomes $S + k_B \ln 2$, and the total entropy becomes $2NS + 2Nk_B \ln 2$, in direct contradiction with the additivity of entropy.

MOMA uses these illicit ensembles, and needs to use them. It has to claim that the volume in state space of a memory cell is twice as big before erasure as it is after erasure. But this only seems to be the case if we use the illicit ensemble of 'random data', the distribution function of which does not correspond to the real thermodynamical state of the individual memory cell. If we think carefully, we must recognise that before erasure the memory is one single state, and after it is still in one single state, with exactly the same accessible volume in state space. Quoting Norton:

> Prior to erasure, the memory device is in state `ZERO` or it is in state `ONE` (but not both!). After the erasure it is in state `ZERO`. Since the states `ZERO` and `ONE` have the same volumes in phase space, there is no change in phase space volume as a result of the erasure procedure. *A process of erasure that resets a memory device in state `ZERO` or in state `ONE` back to the default state `ZERO` does not reduce phase space volume in the sense relevant to the generation of thermodynamic entropy!*[10]

[10] [23], p. 24; emphasis in the original.

7.4.2 From MOMA to SSC

I think it is time to diagnose the sources of the controversy and find out whether Norton's criticism is correct. The main source is the use of different notions of 'ensemble' by different authors.

1. John Norton sees an ensemble as a set of systems which are identical in the sense that they have the same time-averaged behaviour. An acceptable ensemble is one in which the fraction of systems in a part P of state space is equal to the fraction of the time that the systems spend in part P. With this interpretation, there are no ensembles containing both memory cells in state `ZERO` and memory cells in state `ONE`, because these do not have the same time-averaged behaviour. *MOMA* is incoherent.

2. Bennett, Bub, and other exorcists see the ensemble as a measure of ignorance: only if we do *not know* what state the memory cell is in, may we use the ensemble containing memory cells in both states. In this interpretation, the ensemble is not illicit, but has a very clear interpretation. Unfortunately, as Norton points out, it is very hard to see why the 'entropy' thus defined would have anything to do with thermodynamics. *MOMA* is coherent, but not valid.

3. If we interpret the ensemble as I suggested we do in section 2.4, it functions as an indication of the class of systems we wish the demon to be able to operate on. We need to use such an ensemble not as a matter of hard logic, but because we do not wish our demon's success to be excessively sensitive to initial conditions of the system it encounters. With this interpretation the ensemble used by *MOMA* is not illicit, and can be given thermodynamical relevance in the way shown in chapter 3. *MOMA* becomes *SSC*, which is coherent and valid – but makes questionable presuppositions, as shown in chapters 3 and 4.

I conclude that Norton's criticism of *MOMA* is correct insofar as many presentations by exorcists are concerned. But there is an interpretation of the mixed ensemble which does *not* make it illicit, and does preserve the force of *MOMA* by transforming it into *SSC*. Norton's criticism is partly justified, but not entirely correct.

7.5 The new paradigm's failure

Is the application of Landauer's principle the *sine qua non* of successfully exorcising Maxwell's Demon? This chapter leaves little doubt that the answer to this question must be negative. We have seen that there are two classes of proofs of the validity of Landauer's principle: those which make the extended fundamental assumption, and are thus sound but not profound; and those which try to base Landauer's principle on deeper grounds. These latter were either incoherent or, where sense could be made of them, versions of *SSC* embellished with the words 'information' and 'erasure'. And not only could *SSC* be formulated without using those notions or Landauer's principle, it was also seen to be an argument of little value for exorcists.

The erasure-paradigm of exorcism is, then, not the most exalted of stages in the history of mankind's understanding of Maxwell's Demon. Neither does it contain the holy symbols or powerful enchantments needed to banish the demon forever from this world. This does not mean it is wholly a failure: it's thermodynamical proofs, though not profound, nevertheless give insight into the thermodynamics of computation; and arguments like $MOMA$ draw interesting parallels between the thermodynamics of computation and the question of the second law's contingency – even if they are, in and of themselves, not strong enough to answer that question. We would do well to learn from Landauer, Bennett and their followers what we can, and then carefully unlearn the confusion and false beliefs that may have come with that wisdom.

Epilogue: Return of the Demon

Saint Peter's chambers in the Celestial Office. The Demon stands in eager anticipation before a magnificent oaken desk behind which Peter is seated. In the background, we see a gigantic door adorned with a sturdy golden lock.

Saint Peter Well, that was quite a story, little demon. I find it remarkable that you have not given up on the whole project, considering all those scathing *ad hominem* attacks you suffered.

Demon Demons are strong and tough, and prideful besides. Giving up would be an unbearable humiliation.

Saint Peter Yes, of course – and you know where *that* attitude has brought your kind! "To bow and sue for grace, with suppliant knee, and deify his power; that were an ignominy and shame beneath this downfall," or some such rant. I notice that you're back in Heaven anyway, which must have involved *some* kind of groveling!

Demon Looking slightly ashamed. "It's better to reign in Hell than serve in Heaven", yes, yes, don't speak to me about it. Very little reigning done in Hell by the likes of me, I can assure you. Nevertheless, I want to be neither in Heaven nor in Hell – I want to go back to Earth. Has my story convinced you?

Saint Peter The Celestial Office has heard your case, and has reached a decision concerning your request. Takes a formal looking document from the middle of a huge pile of identical-looking forms. The exorcisms which we examined have too many loopholes to hold out against the strength of your determination. If you wish, you may return to Earth.

Demon Great! You've done me a real favour, Pete!

Saint Peter Don't expect a warm welcome, though. I'm not sure they really want you back, down there.

Demon Winks. Then I will *make* them want me. Earth, I'm back!

Saint Peter grabs a huge key and opens the gate. Cackling with only slightly malicious laughter, the Demon jumps through it.

Conclusion

The tale of Maxwell's Demon is a long and complex one, of which only a fraction has been told in this thesis. But from this fraction, several morals have been drawn. It is time to reiterate the main conclusions once more, and thus bring our demonic story to an end.

A being is a successful Maxwell's Demon if it can create an anti-entropic phenomenon with high probability and without endangering its own continued operation. The non-existence of Maxwell's Demon cannot be deduced from classical mechanics alone. Considerations of scales were seen to be very important both in defining what the second law means within the context of classical mechanics, and in showing that classical mechanics cannot, in and of itself, prove the second law. Maxwell's Demon remains possible as long as it has not been assumed that there is a lowest significant scale in nature.

All successful attempts at exorcism actually make this assumption, albeit in the guise of the idea that every system must be described by a canonical distribution function at temperature T – what I call the extended fundamental assumption. The full significance of this assumption has not been appreciated by the exorcists using it. The fluctuation-account of exorcism is sound, but based on this assumption. The measurement-account can be seen as an insightful extension of the fluctuation-account; as a fundamental theory postulating a deep link between entropy generation and measurement, it fails. In the same vein, the erasure-account using Landauer's principle has a sound and an unsound side. If based on the extended fundamental assumption, it provides a correct if limited exorcism. But claims that there is an interesting link between information and entropy on a deeper level do not hold up under scrutiny.

It is very hard to prove the second law on general principles, and certainly it has not yet been done. This recognition is enough to deny the exorcists the full victory they have ofttimes claimed. We must echo the words of Earman and Norton:[11] *The Demon lives*!

[11] [11], p. 25.

Appendix A

Towards a generalised second law

In this appendix, I will take a suggestion from subsection 4.2.4 and indicate how a variant of the fine-grained Gibbs entropy can be used to formulate a generalised second law. The aim is to make explicit considerations of scale, and arrive at a second law which is supposed to hold independent of the contingent facts identified in part I, and still be useful to physicists. I succeed only partly, which explains the 'towards' in the title.

Let $\Omega = \{\ldots, \Omega_{-1}, \Omega_0, \Omega_1, \ldots\}$ be a set of sets, which designates a system. The Ω_n are sets of objects of a scale n, where higher n denotes bigger objects. Thus,

$$\Omega_n = \{O_{n,1}, O_{n,2}, \ldots, O_{n,m(n)}\}, \tag{A.1}$$

where $m(n)$ is the number of objects of scale n, and $O_{i,j}$ is the j-th object on scale i. Let $K(O_{i,j})$ be the kinetic energy of object $O_{i,j}$, calculated in the frame of rest of the bigger object it is a part of.[1] Then

$$K_n = \sum_{j=1}^{m(n)} K(O_{n,j}) \tag{A.2}$$

is the total kinetic energy on scale n. Furthermore, P is the combined potential energy of all objects – a scale-transcending notion – so that the total energy is

$$E = P + \sum_{n=-\infty}^{\infty} K_n. \tag{A.3}$$

The first law of thermodynamics is then:

$$\Delta E = 0. \tag{A.4}$$

We now define a state space for every scale, which is spanned by the position and momentum vectors of every object on that scale. Thus

$$\Gamma_n = \bigoplus_{i=1}^{m(n)} (\vec{q}_{O_{n,i}}, \vec{p}_{O_{n,i}}). \tag{A.5}$$

[1] See subsection 3.1.1

The total state space Γ is defined as

$$\Gamma = \bigoplus_{n=-\infty}^{\infty} \Gamma_n. \tag{A.6}$$

Let $\rho(\vec{x})$ be a normalised distribution function on Γ, with the additional property that it can be decomposed as follows:[2]

$$\rho(\vec{x}) = \prod_{n=-\infty}^{\infty} \rho_n(\vec{x}_n), \tag{A.7}$$

where ρ_n is a normalised distribution function on Γ_n. We define the **entropy on scale** n analogously to the fine-grained Gibbs entropy of formula 2.2:

$$S_n = S_n[\rho_n(\vec{x}_n)] = -k_B \int_{\Gamma_n} \rho_n(\vec{x}_n) \ln\{\rho_n(\vec{x}_n)\} d\vec{x}_n. \tag{A.8}$$

Additionally, we define the **total entropy** as:

$$S = S[\rho(\vec{x})] = -k_B \int_{\Gamma} \rho(\vec{x}) \ln\{\rho(\vec{x})\} d\vec{x}. \tag{A.9}$$

It can easily be shown that the total entropy is equal to the sum of the scale dependent entropies:

$$\begin{aligned}
S &= -k_B \int_{\Gamma} \rho(\vec{x}) \ln\{\rho(\vec{x})\} d\vec{x} && \text{(A.10)}\\
&= -k_B \int_{\Gamma} \{\prod_{n=-\infty}^{\infty} \rho_n(\vec{x}_n)\} \ln\{\prod_{n=-\infty}^{\infty} \rho_n(\vec{x}_n)\} d\vec{x} && \text{(A.11)}\\
&= \sum_{i=-\infty}^{\infty} -k_B \int_{\Gamma} \{\prod_{n=-\infty}^{\infty} \rho_n(\vec{x}_n)\} \ln\{\rho_i(\vec{x}_i)\} d\vec{x} && \text{(A.12)}\\
&= \sum_{i=-\infty}^{\infty} -k_B \int_{\Gamma/\Gamma_i} \{\prod_{n\in\{\mathbb{N}/i\}} \rho_n(\vec{x}_n)\} d\vec{x} \times && \text{(A.13)}\\
&\qquad \int_{\Gamma_i} \rho_i(\vec{x}_i) \ln\{\rho_i(\vec{x}_i)\} d\vec{x}_i && \text{(A.14)}\\
&= \sum_{i=-\infty}^{\infty} -k_B \int_{\Gamma_i} \rho_i(\vec{x}_i) \ln\{\rho_i(\vec{x}_i)\} && \text{(A.15)}\\
&= \sum_{i=-\infty}^{\infty} S_i. && \text{(A.16)}
\end{aligned}$$

We can follow $\rho(\vec{x})$ through time, giving us the function $\rho(\vec{x}, t)$. Assume that this can still be decomposed into now time-dependent functions on the subspaces of Γ which correspond to different scales:

$$\rho(\vec{x}, t) = \prod_{n=-\infty}^{\infty} \rho_n(\vec{x}_n, t). \tag{A.17}$$

[2]This is a very strong, but not a very crucial, assumption. We'll come back to it.

We now have the tools to formulate a **generalised second law of thermodynamics**:

Suppose D is a machine which, operating in a cycle on a system Ω, changes every cycle with high probability the K_n in the following way:

$$\sum_{n=-\infty}^{a} K_n^I < \sum_{n=-\infty}^{a} K_n^F, \tag{A.18}$$

where F stands for 'final', and I for 'initial'; and furthermore

$$\Delta K_a \equiv K_a^F - K_a^I < 0. \tag{A.19}$$

Then with high probability

$$\sum_{n=-\infty}^{a-1} \Delta S_n \approx -\Delta S_a > 0. \tag{A.20}$$

If there is a smallest scale c, so that all Γ_i with $i < c$ are empty and all S_i with $i < c$ are zero, this condition cannot be met. As a consequence, if there is a smallest scale, no machine can exist which with high probability, operating in a cycle, can transform the kinetic energy of objects on this smallest scale to energy on higher scales or potential energy. Given the contingent fact that there are no tinyons, this reduces to the normal statistical version of the second law of thermodynamics.

The idea behind the above formulation of the laws is as follows. Suppose we make a machine which operates in a cycle and succeeds with high probability in changing the kinetic energy of random motion on a certain scale a to kinetic energy on a larger scale, or to potential energy. This machine would be a successful Maxwell's Demon at scale a. What the generalised second law claims is that this Maxwell's Demon *must* increase the entropy (diversity of the ensemble) at one or more scales beneath a. And as a consequence, if there are no scales beneath a, there can be no Maxwell's Demon at scale a.

It is important to notice the somewhat bizarre way in which we have used the fine-grained Gibbs entropy in this law. The total entropy of the system, S, remains constant if we allow only Hamiltonian time-evolutions – this was the great problem of this measure of entropy. But the scale-dependent entropies S_n do *not* have to remain constant. They measure the 'diversity of the ensemble at a certain scale', but this diversity can be transformed from one scale to the other. For instance, in normal dissipation processes both energy and 'diversity' is transferred from a higher to a lower scale. And as we have seen in chapters 3 and 4, because there are so many fewer degrees of freedom at higher scales than at lower scales, diversity cannot be transferred to higher scales, but can be transferred to lower one.

How does one go about proving this generalised second law? Obviously, $\rho(\vec{x}, 0)$ is to be interpreted as the initial ensemble of systems. Then, Hamiltonianism gives us that $\Delta S = 0$, and thus that the sum of the S_n is constant. Next, we must prove that if we exploit random motion of objects on scale a,

then $\Delta K_a < 0 \rightarrow \Delta S_a < 0$. Such an argument is produced in section 3.2.1, where it is shown that volume in state space increase very fast with available kinetic energy. Of course in general a change in kinetic energy says nothing about volumes in state space (for instance, if you have only one degree of freedom, your kinetic energy can change but your volume will not), but for general, 'random' ensembles it does. A realistic ensemble for a hot gas will be spread out across a much larger part of state space than that for a cold gas. We may now conclude that if kinetic energy on scale a is transformed into 'higher' forms of energy, $\Delta S_a < 0$. Consequently, the sum over all other S_n must become bigger. The generalised second law claims that at least part of this increase must be at scales smaller than a. The elements needed for a proof of this further include a demonstration, akin to that given in section 3.2.1, that the fraction $\Delta S_n / \Delta K_n$ increases very rapidly with the number of relevant degrees of freedom on scale n, which is $3m(n)$. If we then put in the assumption that $m(a) \gg \sum_{n=a+1}^{\infty} m(n)$, the generalised second law will follow. For evidently, the loss of kinetic energy at scale a implies a loss of entropy which cannot be balanced by increases of kinetic energy at higher scales. Ergo, it must be balanced by increases of energy at lower scales.

The present law has a few weaknesses which make it somewhat less than a successful generalised second law. The first thing that will be noticed by readers are the assumptions A.7 and A.17. It is not in general very plausible that the ensemble $\rho(\vec{x})$ can be so decomposed, and it is certainly quite implausible that it will remain in such a mathematical form under evolution. This might not be a deep problem, as we can opt to define

$$\rho_i(\vec{x}_i) = \int_{\Gamma/\Gamma_i} \rho(\vec{x}) d\vec{x}_{\mathbb{N}/i}, \tag{A.21}$$

after which it may still be possible to prove that the total entropy is equal to the sum of the scale-dependent entropies. I confess that I have not looked into this mathematical problem.

A more fundamental weakness is that the law still needs two major assumptions, contingent given only classical mechanics, in order to be proved at all. The first is that the world can be divided into scales. If this is not granted, the whole formalism used just collapses. The second major assumption is that $m(a) \gg \sum_{n=a+1}^{\infty} m(n)$, that there are much more objects (or degrees of freedom) at scale a than at all higher scales combined. This furnishes the scale-asymmetry which we absolutely need to get the second law of the ground in any form like its original one. Yet a third major assumption, that $m(a) \gg \sum_{n=-\infty}^{a-1} m(n)$, is needed in order to have the normal second law follow from the 'generalised' one.

The generalised law here developed is still a contingent law, given only classical mechanics. By adding the two major assumptions noted above as 'if'-clauses to the law, it would become necessary, a consequence of classical mechanics and considerations of ensembles and probabilities and such; but its connection with the original second law would be very vague at best, its usefulness for thermal physics negligible. There is a trade-off, a delicate balance, between usefulness and lack of assumptions. Thermal physicists in the past have clearly opted to put more weight on the former, and I tend to agree with them. Generalising the second law may well turn out to be a pass-time for philosophers, of which all true physicists stay clear.

Bibliography

[1] **Albert, David Z.**: *Time and Chance*, Harvard University Press (2000)

[2] **Bennett, C.H.**: *The thermodynamics of computation – a review*, International Journal of Theoretical Physics, Vol. 21, p. 905-940 (1982); included in Leff & Rex 2003, p. 283-318.

[3] **Bennett, C.H.**: *Demons, Engines, and the Second Law*, Scientific American, Vol. 257:5, p. 108-116 (1987).

[4] **Bennett, C.H.**: *Notes on the history of reversible computation*, IBM Journal of Research and Development, Vol. 32, p. 16-23 (1988).

[5] **Bennett, C.H.**: *Information physics in cartoons*, Superlattices and microstructures, Vol. 23, p. 367-372 (1998).

[6] **Bennett, C.H.**: *Notes on Landauer's principle, reversible computation, and Maxwell's Demon*, Studies in the History and Philosophy of Modern Physics, Vol. 34B, No. 3, 501-510 (2003)

[7] **Brillouin, L.**: *Maxwell's demon cannot operate: Information and entropy. I*, Journal of Applied Physics, Vol. 22, 338-343 (1951)

[8] **Brillouin, L.**: *Science and Information Theory*, Second edition, Academic Press Inc. (1962)

[9] **Bub, Jeffrey**: *Maxwell's Demon and the Thermodynamics of Computation*, arXiv:quant-ph/0203017 v1 (2002)

[10] **Earman, John & John D. Norton**: *Exorcist XIV: The Wrath of Maxwell's Demon. Part I. From Maxwell to Szilard*, Studies in the History and Philosophy of Modern Physics, Vol. 29, No. 4, 435-471 (1998)

[11] **Earman, John & John D. Norton**: *Exorcist XIV: The Wrath of Maxwell's Demon. Part II. From Szilard to Landauer and Beyond*, Studies in the History and Philosophy of Modern Physics, Vol. 30, No. 1, 1-40 (1999)

[12] **Feynman, Richard P., Robert B. Leighton & Matthew Sands**: *The Feynman Lectures on Physics*, Volume 1, Addison-Wesley Publishing Company (1963)

[13] **Garber, Elizabeth; Stephen G. Brush & C.W.F. Everitt**: *Maxwell on Molecules and Gases* (The MIT Press, 1986)

[14] **Garber, Elizabeth; Stephen G. Brush & C.W.F. Everitt**: *Maxwell on Heat and Statistical Mechanics. On "Avoiding All Personal Enquiries" of Molecules* (Associated University Presses, Inc., 1995)

[15] **Heiman, P.M.**: *Molecular forces, statistical representation and Maxwell's demon*, Studies in the History and Philosophy of Science, Vol. 1, 189-211 (1970); reprinted in [19], p. 52-74

[16] **Horwich, P.**: *Asymmetries in Time: Problems in the Philosophy of Science*, MIT Press (1987)

[17] **Landauer, R.**: *Irreversibility and heat generation in the computing process*, IBM Journal of Research and Development, Vol. 5, p. 183-191 (1961)

[18] **Lavis, David**: *Some Examples of Simple Systems*, version2, www.mth.kcl.ac.uk/˜ dlavis/papers/examples-ver2.pdf (2003)

[19] **Leff, Harvey S. & Andrew F. Rex**: *Maxwell's Demon: entropy, information, computing* (Institute of Physics, 1990)

[20] **Leff, Harvey S. & Andrew F. Rex**: *Maxwell's Demon 2: entropy, classical and quantum information, computing* (Institute of Physics, 2003)

[21] **Magnasco, Marcelo O. & Gustavo Stolovitzky**: *Feynman's Ratchet and Pawl*, Journal of Statistical Physics, Vol. 93, p. 615-632 (1998)

[22] **Maxwell, James Clerk**: *Diffusion*, Encyclopedia Brittanica, ninth edition, Vol. 7, p. 214-221 (1878); reprinted in [13], p. 524-546

[23] **Norton, John D.**: *Eaters of the Lotus: Landauer's Principle and the Return of Maxwell's Demon*, PhilSci Archive, http://philsci-archive.pitt.edu/archive/00001729/ (2004)

[24] **Piechocinska, Barbara**: *Information erasure*, Physical Review A, Vol. 61 (2000)

[25] **Reichenbach, Hans**: *The Direction of Time*, University of Los Angeles Press (1956)

[26] **Shannon, C.E.**: *A Mathematical Theory of Communication*, The Bell System Technical Journal, Vol. 27, p. 379-423 (1948)

[27] **Shizume, K.**: *Heat generation required by erasure*, Physical Review E, Vol. 52, p. 3495-3499 (1995)

[28] **Sklar, Lawrence**: *Physics and Chance. Philosophical Issues in the Foundations of Statistical Mechanics*, (Cambridge University Press, 1993)

[29] **Skordos, P.A. & W.H. Zurek**: *Maxwell's demon, rectifiers, and the second law: Computer simulation of Smoluchowski's trapdoor*, American Journal of Physics, Vol. 60, No. 10, p. 876-882 (1992)

[30] **Smoluchowski, M. v.**: *Experimentell nachweisbare, der üblichen Thermodynamik widersprechende Molekularphänomene*, Physikalische Zeitschrift, vol. 13, p. 1069-1080 (1912)

[31] **Smoluchowski, M. v.**: *Gültigkeitsgrenzen des zweiten Hauptsatzes der Wärmtetheorie*, in 'Vorträge über die kinetische Theorie der Materie und der Elektrizität', Teubner (1914)

[32] **Szilard, Leo**: *Über die Entropieverminderung in einem thermodynamischen System bei Eingriffen intelligenter Wesen*, Zeitschrift für Physik, vol. 53, p. 840-856 (1929); English translation *On the decrease of entropy in a thermodynamic system by the intervention of intelligent beings*, in Leff & Rex (1990), [19], p. 124-133

[33] **Zhang, Kechen & Kezhao Zhang**: *Mechanical models of Maxwell's demon with noninvariant phase volume*, Physical Review A, Vol. 46, No. 8, p. 4598-4605 (1992)