DAVID Z ALBERT

DAVID Z ALBERT

TIME AND CHANCE

TIME AND CHANCE

TIME AND CHANCE

DAVID Z ALBERT

TIME AND CHANCE

HARVARD UNIVERSITY PRESS
Cambridge, Massachusetts
London, England

Printed in the United States of America
Second printing, 2003

Library of Congress Cataloging-in-Publication Data

Albert, David Z.
Time and chance / David Z Albert.
p. cm.
Includes index.
ISBN 0-674-00317-9 (cloth)
ISBN 0-674-01132-5 (paper)
1. Time reversal. 2. Physics—Philosophy. I. Title.

QC173.59.T53 A43 2000
530.11—dc21 00-056732

FOR BEN

CONTENTS

PREFACE

This book is intended both as an elementary introduction and as an original contribution to the development of a scientific account of the distinction between the past and the future.

▲▲▲ Chapter 1 is a relatively straightforward rehearsal of what is perennially referred to in the physical literature as "the problem of the direction of time"—but in what I hope is an unprecedentedly precise language of physical states, and with what I hope is an unprecedentedly careful discussion of exactly what it means for a set of dynamical laws to distinguish, or to *fail* to distinguish, between the past and the future. Chapter 2, together with the more detailed appendix on Gedankenexperiments with heat engines, is a report on the second law of thermodynamics—which is the point at which distinctions between past and future have made their most explicit and most widely heralded and most intensively studied appearance in the laws of physics—very much along the lines of the beautiful treatment of that subject in the famous book by Enrico Fermi. Chapter 3 is an outline, more or less in the spirit of Ludwig Boltzmann, of the project of statistical mechanics—including what I hope will prove relatively novel discussions of the mathematical structure and the metaphysical status of the probability-distributions over initial conditions, and of the connection between entropy and information, and of the question of Haecceisstism, and of a number of other matters as well. Chapter 4 is about the famous objections to that project due to J. Loschmidt and Ernst Zermello and Henri Poincaré, and also about what seems to me to be the proper *remedy* for those objections—which is a new and fundamental and non-dynamical law of nature called the "past-hypothesis." Chapter 5 is a critique of the long history of at-

tempts to show that there can as a matter of fundamental principle be no such thing as an operational Maxwellian demon. What I argue (on the contrary—and by means of an explicit construction) is that there is nothing whatsoever in either the classical or the quantum-mechanical laws of physics that stands in the way of there *being* such demons as that. Chapter 6 (which is perhaps the most ambitious) is about the time-directedness of our own capacity to *acquire information* about the world, and to *manipulate* the world according to our will, or (more precisely) it is about the business of *incorporating* those sorts of directedness into the general picture of the world laid out thus far, that is, it is about the business of *understanding* those sorts of directedness as *mechanical phenomena of nature*—of a piece (that is) with the understanding of the time-directedness of the *second law of thermodynamics* described in chapters 3 and 4. And Chapter 7 (which is perhaps the most *important*) is about the possibility of there being a very illuminating connection between the problem of the direction of time and the fundamental quantum-mechanical problem of *measurement*.

▲▲▲ There are so many people to thank. I am thankful first and foremost and beyond all reckoning to my wife, Orna, for the weeks and months and years of hard work she did so that I might have time to write, and for her patience and her encouragement, and for the unwavering enormity of her love, and for figuring out that what this sort of book needed on its cover was Italian Futurism. And I am thankful to Tim Maudlin and Frank Artzenius and Ned Hall and Simon Saunders and Gordon Belot and especially and particularly to Barry Loewer for their extraordinarily careful readings of my manuscript—in the course of which many errors (both small and large) were corrected. And to Bas van Fraassen and David Lewis and various of my students at Columbia and Princeton and Harvard who have patiently heard me out on these matters in seminars and lecture courses, and have helped me get them straight. And to Yakir Aharonov and Hilary Putnam and Marc Albert and Arthur Kantrowitz and the late Gary Feinberg for whatever I may have managed to absorb of the general craft of thinking things through. And to Sidney Morgenbesser and Irad Kimhi, of whom (let's just say) indescribably vast stretches of my work, and of my *self*, are poor attempts at imitation. And I am thankful to my friend Lloyd Miller for preparing the diagrams, and to Lindsay Waters and Kim Steere and Christine Thorsteinsson and

Elizabeth Gilbert and Jill Breitbarth of Harvard University Press for editing and designing and publishing my manuscript with such a profound and imaginative respect for my tone of voice (such as it is) and for the sort of book I've tried to write.

TIME AND CHANCE

TIME-REVERSAL INVARIANCE

1. THE NEWTONIAN PICTURE OF THE WORLD

What I want to talk about here is a certain tension between fundamental microscopic physical theory and everyday macroscopic human experience, a tension that comes up (more particularly) in connection with the question of precisely how the past is different from the future.

And the fundamental theory in which it will work best to start that talk out, the fundamental theory (that is) in which this tension is at its purest and most straightforward, is the mechanics of Newton. Never mind (for the moment) that the mechanics of Newton turns out not to be the mechanics of the actual *world*.[1] We'll talk about that later.

▲▲▲ According to Newtonian mechanics, or at any rate according to the particularly clean and simple version of it that I want to start off with here, the physical furniture of the universe consists entirely of point particles. The only *dynamical variables* of such particles—the only physical attributes of such particles that can *change with time*—are (on this theory) their *positions;* and (consequently) a list of what particles exist, and of what *sorts* of particles they are,[2] and of what their positions are at all times, is a list of absolutely everything there is to say about the physical history of the world.[3]

1. What the actual *world* turns out to be (insofar as we can tell at present) is *quantum* mechanical, or quantum field theoretic, or quantum string theoretic, or something like that.

2. That is, of their *non*-dynamical properties: their *masses* and their *charges* and so forth.

3. This is certainly not to deny that there are such things in the world as *extended objects;* the idea is just that all the facts about objects like that (facts, say, about where the tables and chairs are, and about who punched whom, and about who said what, and so on) are determined, in principle, by the facts about the *particles* of which those objects are *composed*.

And Newtonian mechanics is *deterministic*. Given a list of the positions of all the particles in the world at any particular time, and of how those positions are *changing*, at that time, as time flows forward, and of what *sorts* of particles they are, the universe's entire history, in every detail, from that time on, can in principle be calculated (if this theory is true) with certainty.

▲▲▲ All this will be worth spelling out in some detail.

Let's start slow.

The rate at which some particular particle's position is changing, at some particular time, as time flows forward, is called its *velocity* at that time. And the rate at which such a particle's *velocity* is changing, at some particular time, as time flows forward, is called its *acceleration* at that time.

And what Newtonian mechanics has to say about the motions of particles, the *entirety* of what it has to say about the motions of particles, is that a certain breathtakingly simple mathematical relation—$F = ma$—invariably holds between the force on any particle at any particular instant, and its acceleration at that instant, and its mass.

Let's say a bit about where forces come from.

What happens in the most familiar cases (think, for example, of gravitational attraction, or of electrostatic repulsion) is that forces arise exclusively *between pairs of particles*, and (moreover) that the forces which any two particles are exerting on each other at any particular instant depend only on *what sorts of particles they are* and on their relative *positions*.

And the third and final fundamental principle of the Newtonian picture of the world (the first is that the world consists entirely of particles, and the second is the relation between F and m and a) is that as a matter of fact all the forces there *are* are like that.

And so (on this picture) a specification of the positions of all the particles in the world at some particular time, and of what sorts of particles they are, amounts (at least insofar as these familiar sorts of forces are concerned) to a specification of what the forces on each of those particles are at that time as well.

Good. Let's see how all this results in precisely the sort of determinism I said it did above.

Call the "initial" time, the time we will want to calculate forward from, $t = 0$.

And suppose that what we're given at the outset are the positions of all the particles in the world (or in some isolated subsystem of the world) at $t = 0$ (call those x_0^i), and their velocities at $t = 0$ (call those v_0^i), and their masses (m^i), and their electric charges (c^i), and all their other intrinsic properties.

And let's say that what we would like to calculate is the positions of all these particles at $t = T$.

The most illuminating way of doing that, I think, will be by means of a succession of progressively better and better *approximations*.

The first goes like this: calculate the positions of all the particles at $t = T$ by supposing that their velocities are constant—and equal to their above-mentioned values (v_0^i) at $t = 0$—throughout the interval between $t = 0$ and $t = T$.

This calculation will place particle i at $x_0^i + v_0^i T$ at $t = T$; but it hardly needs saying that this calculation is not a particularly accurate one, because (unless it happens that no forces are at work on any of the particles here) the velocities of these particles will in fact *not* remain constant throughout that interval.

Here's a somewhat better calculation:

Divide the time-interval in question into two, one extending from $t = 0$ to $t = T/2$ and the other extending from $t = T/2$ to $t = T$. Then calculate the positions of all the particles at $T/2$ by supposing that their velocities are constant—and equal to their values at $t = 0$—throughout the interval between $t = 0$ and $t = T/2$ (this will place particle i at $x_0^i + v_0^i(T/2)$ at $T/2$).

Then calculate the *forces* acting on each of the particles at $t = 0$ (what those forces are, remember, will follow from the *positions* of those particles at $t = 0$ together with their masses and their charges and their other internal properties—all of which we are given at the outset).

Then calculate each particle's *velocity* at $T/2$ by plugging those forces into the above-mentioned law of motion (plugging them, that is, into $F = ma$), and assuming that the particles' accelerations are constant throughout the interval from $t = 0$ to $t = T/2$—and are equal to their values at $t = 0$ (this will

put the velocity of particle i at $v_0^i + a_0^i(T/2)$, where a_0^i is equal to the force on particle i at $t = 0$ divided by particle i's mass).

Then, finally, calculate the position of particle i at $t = T$ (which is what we're after here) by supposing that this particle maintains this *new* velocity throughout the interval between $t = T/2$ and $t = T$.

This calculation isn't going to be perfect either, but (since the intervals during which the velocities of the particles are erroneously presumed to be constant are shorter here than in the previous calculation) it amounts to a clear improvement.

And of course this improvement can *itself* be improved upon by dividing the interval further, into *four* intervals. That calculation will proceed as follows.

To begin with, the approximate positions of all the particles at $t = T/4$ can be calculated (just as we did above) from the positions and velocities at $t = 0$ alone. Moreover, the forces on all the particles at $t = 0$ can now be read off (as we did above) from their intrinsic properties and their positions at $t = 0$, and thus (with the aid of $F = ma$) the approximate *velocities* of all the particles at $t = T/4$ can be deduced as well. And so what we now have in hand is a list of approximate positions and approximate velocities and particle-types at $T/4$, and of course those approximate positions and particle-types can *now* be used to determine the approximate *forces* on all the particles at that time as well, and *that* will in turn allow us to determine the positions and velocities and forces at $T/2$, and so on.

And then we can go on to eight intervals, and then to sixteen. And as the number of intervals approaches *infinity*, the calculation of the particles' positions at $t = T$ patently approaches perfection. And it happens that there is a trick (and the name of that trick is *the calculus*) whereby—given a simple enough specification of the dependence of the forces to which the particles are subjected on their relative positions—that perfect calculation can actually and straightforwardly be carried out.

And of course T can be chosen to have any positive value we like. And so the positions of all the particles in the system in question at any time between $t = 0$ and $t =$ infinity (and with that the *velocities* of all those particles between those times, and their energies, and their angular momenta, and everything else about them) can in principle be calculated, exactly and with

certainty, from the positions and velocities and intrinsic properties of all those particles at $t = 0$.[4]

2. TIME-REVERSAL INVARIANCE IN THE NEWTONIAN PICTURE

Newtonian mechanics has a number of what are referred to in the literature as *fundamental symmetries;* and what that means is that in Newtonian mechanics there are certain sorts of facts about the world which—as a matter of absolutely general principle—*don't make any dynamical difference.*

Suppose, for example, that we are given the present positions and velocities of all the particles in the world, and that we are told what *sorts* of particles they are, and that we would like to calculate their positions and velocities (say) two hours from now. It is an extremely straightforward consequence of

4. Note, by the way, that this *overall* determinism of the evolution of a universe of classical particles (whereby all the present positions and velocities determine all the *future* ones) can invariably be *taken apart* (as it were) into *separate* determinisms running in *parallel.*

Consider, for example, a universe consisting of a single classical particle, which (being all alone in the world) is never subjected to force.

Exactly two logically independent pieces of information—two *numbers*—are required in order to specify fully the present physical situation of such a particle, and to nail down (by means of the classical laws of motion) all its future situations. The pair of numbers we're *usually* presented with in contexts like this is the particle's present *position* (x_0) and its present *velocity* (v_0); but there are, of course, an infinite collection of *other, mathematically equivalent,* such pairs ($x_0 + v_0$ and $x_0 - v_0$, for example, or $5x_0 + 14v_0$ and $36x_0 - 7v_0$, or x_0 and $x_0 + 23v_0$, or what have you) which will patently do just as well.

Imagine, then, that we are informed of the values of the quantities v_0 and $x_0 + Tv_0$ (where T is some number), and that we would like to deduce, by means of the laws of classical mechanics, the values of the velocity and the position of the particle at some later time $t = T$. The calculations involved here are trivial, of course, but what I want to draw the reader's attention to here is that the outcome of the first of those calculations (which is: $v_T = v_0$) will depend *exclusively* on the value of v_0 and *not at all* on the value of $x_0 + Tv_0$, and that the outcome of the *second* of those calculations (which is: $x_T = x_0 + Tv_0$) will depend exclusively on the value of $x_0 + Tv_0$ and not at all on the value of x_0. And so if we were informed only of the value of v_0, and were left entirely in the dark about $x_0 + Tv_0$, we could nonetheless deduce, with certainty, from the laws of classical mechanics alone, the value of v_T; and if we were informed only of the value of $x_0 + Tv_0$, and were left entirely in the dark about v_0, we could nonetheless deduce, with certainty, from the laws of classical mechanics alone, the value of x_T.

And this turns out to be an absolutely general phenomenon, which applies to classical worlds consisting of any number of particles, interacting with one another in any way you like: the velocity (say) of particle number 789 at $t = 6$ years will necessarily be equal to some definite function of the positions and velocities of all the particles in the world at $t = 0$, and the position of particle number 3 at that time will necessarily be some *other* such function, and the values of those two functions will necessarily be logically *independent* of each other.

the Newtonian picture of the world I described above that that calculation can be carried through in perfect ignorance of what time "now" is. If the classical laws of motion entail that a certain set of positions and velocities at 4:02 evolves into a certain *other* set of positions and velocities at 6:02, then those laws will necessarily *also* entail that the first set at 4:07 will evolve into the second set at 6:07, and that the first set at 12:23 will evolve into the second at 2:23, and so on. Any sequence of position and velocity values for every particle in an isolated collection which is in accord with classical mechanics and which begins at time t would necessarily (to put all this slightly differently) *also* be in accord with classical mechanics if it were to begin instead at t'. And in virtue of all that, Newtonian mechanics is said to have *time-translation-symmetry*, that is, it is said to be *invariant* under translations like that.

And it has a number of other significant invariances as well: absolute *positions* don't play any role in Newtonian mechanics (although the positions of particles *relative to one another* certainly do), and neither do absolute *directions in space*, and neither do absolute *velocities*.[5]

▲▲▲ And neither does *the direction of time*.

Let's stop and talk about that some.

Imagine, to begin with, watching a film of a baseball which is thrown directly upward, and which is subject to the influence of the gravitational force of the earth; and then imagine watching the same film *in reverse*. The film run forward will depict the baseball moving more and more slowly upward; and the film run in reverse will depict the baseball moving more and more quickly downward. What *both* films will depict, though, is a baseball which (whatever its *velocity*) is accelerating, constantly, at the rate of 32 feet per second per second, *in the direction of the ground*.

And this is (of course) an absolutely general phenomenon: the apparent velocity of any particular material particle at any particular frame of any film of any classical physical process run forward will be equal and opposite to the apparent velocity of that particle at that frame of that film run in reverse; but the apparent *acceleration* of any particular particle at any particular frame of the film run forward will be *identical*, *both* in magnitude *and* in di-

5. And all of these invariances are (by the way) also invariances of Maxwellian electrodynamics, and of relativistic quantum string theories, and of every other fundamental theory in the canon too.

rection, to the apparent acceleration of that particle at that frame of the film run in reverse.[6]

Now, the Newtonian law of motion (which is, remember, the entirety of what the Newtonian picture of the world has to say about the motions of particles) is that a certain mathematical relation holds, at every instant, between mass and force and acceleration. And of course the mass of any particular particle at any particular frame of the sort of movie we've been talking about depends on nothing other than *what* particular particle it *is;* and the *force* on any particular particle at any particular frame of the sort of movie we've been talking about depends on nothing other than what particular *set* of particles happens to exist, and what those particles' spatial distances from one another at that frame happen to be; and what we've just seen is that the *acceleration* of any particular particle, at any particular frame of such a movie, will be entirely independent of the direction in which the film is run. And so if a certain film, run forward, depicts a process which is in accord with Newtonian mechanics, then, necessarily, the same film run in *reverse* will depict a process which is in accord with Newtonian mechanics as well.[7]

6. The proof is trivial. Let $x(t)$ represent the apparent trajectory (that is, the apparent position as a function of time) of the particle depicted in the film run forward, and let $v(t)$ represent the apparent *velocity* (that is, the apparent *derivative* of the position with *respect* to time) of that particle, and let $a(t)$ represent the apparent *acceleration* (that is, the apparent derivative of the *velocity* with respect to time) of that particle. Then the apparent trajectory of the particle depicted in the film run in *reverse* will be $x(-t)$. And of course the apparent *velocity* of a particle whose apparent trajectory is $x(-t)$ is (by definition) $dx(-t)/dt$, which is equal (by the chain rule) to $-v(-t)$, which is (of course) the negative of the velocity of the particle (at the frame in question) depicted in the film run forward. By contrast, the apparent *acceleration* of a particle whose apparent trajectory is $x(-t)$ is (by definition) $(d/dt)(dx(-t)/dt)$, which is equal (by the upshot of the previous sentence) to $(d/dt)(-v(-t))$, which is equal (by the chain rule) to $-(-a(-t))$, which is equal (because (-1) times (-1) is $(+1)$) to $a(-t)$, which is (of course) the *same* as the acceleration of the particle (at the frame in question) depicted in the film run forward.

7. Let's put this a bit more formally. Consider a history $\{x_1(t) \ldots x_N(t)\}$ of some isolated collection of N particles. What's just been shown is that if

$$d^2x_i(t)/dt^2 = F_i(x_1(t) \ldots x_N(t))$$

for all i (where x_i is the position of particle i, and F_i is the force *on* particle i), then necessarily

$$d^2x_i(-t)/dt^2 = F_i(x_1(-t) \ldots x_N(-t));$$

which is to say that if $\{x_1(t) \ldots x_N(t)\}$ is a solution to the Newtonian equations of motion, then necessarily

$$\{x_1(-t) \ldots x_N(-t)\}$$

is too.

And so it is a consequence of Newtonian mechanics that nothing in the laws of nature can be of any help whatsoever in deciding *which way* any film is ever being run. And so it is a consequence of Newtonian mechanics that whatever can happen can just as easily, just as *naturally*, happen *backward*.[8]

And so the Newtonian-mechanical instructions for calculating future physical situations of the world from its present physical situation turn out to be identical to the Newtonian-mechanical instructions for calculating *past* physical situations of the world from its present physical situation. The instructions for calculating (say) the positions of all the particles in the world ten minutes from now are to plug the present positions of all those particles, and the rates at which those positions are changing as time flows forward, into a certain algorithm (the sort of algorithm we explicitly went through above); and the instructions for calculating the positions of those particles ten minutes *ago* are to plug their present positions, and the rates at which those positions are changing as time flows *backward*, into precisely the *same* algorithm.

And so if we are told the positions of all the particles in the world at present, and if we are told the rates at which those positions are changing as time flows toward some other moment *M*, and if we are told the size of the time-interval that separates *M* from the present, then we can in principle calculate the positions of all the particles in the world at *M*, with certainty, without *ever* having been told (and also without ever *learning*, as the calculation *proceeds*) whether *M* happens to lie *after* the present or *before* it.

And so (if the laws of Newtonian mechanics are all the fundamental natural laws there are) there can be no lawlike asymmetries whatsoever between past and future.

▲▲▲ And the thing is that all this is wildly at odds with our everyday experience.

To begin with, every corner of the world is positively *swarming* with ordinary physical processes that don't, or don't regularly, or don't naturally, or

8. Maybe a few of the standard illustrations are in order here. Think, then, of watching films, run forward and run in reverse, of a single particle, alone in the universe, moving (say) to the right; or of two billiard balls colliding; or of a rock moving downward, and accelerating downward, in the gravitational field of the earth.

don't familiarly, happen backward (the melting of ice, say, or the cooling of soup, or the breaking of glass, or the passing of youth; whatever).[9]

And (on top of that) there's an asymmetry of *epistemic access:* our capacities to know what happened yesterday, and our methods of *finding out* what happened yesterday, are as a general matter very different from our capacities to know and our methods of finding out what will happen *tomorrow*.[10]

And (on top of *that*) there's what you might call an asymmetry of *intervention:* it seems to us that we can bring it about that certain things occur—or that they don't—in the future, but we feel absolutely incapable of doing anything at all about the *past*.

▲▲▲ And that's the tension I mentioned before. And that's more or less what I want to talk about in this book. Or at any rate that's the *Newtonian* version.

The next thing to do is to generalize it some.

3. TIME-REVERSAL INVARIANCE IN GENERAL

Let's start by thinking through what it means to give a complete description of the physical situation of the world at an instant.

There would seem to be two things you want from a description like that:

a. that it be genuinely *instantaneous* (which is to say that descriptions of the world at different times have the appropriate sort of logical or conceptual or metaphysical *independence* of one another, that a perfectly explicit and intelligible *sense* can be attached to *any temporal sequence whatever* of the sorts of descriptions we have in mind here—whether the sequence happens to be in accord with the dy-

9. Maybe this is worth belaboring a bit further. Take soup. It isn't that soup never *heats up*; it's (rather) that occasions when soup *does* heat up never look anything at all like mere *temporal inversions*, like *films watched backward* of occasions when soup *cools off*. The former occasions are always *different*, somehow. They involve fires or electrical currents or parabolic mirrors or something like that. And *that's* the point here: that you can bet your life that a tepid pot of soup, in (say) an otherwise empty, cold, closed, insulated room, is invariably and ineluctably in the process of getting *colder*.

10. Note that this is no less a physical business than the stuff about the cooling of soup: this too is about the sequences in which the *states of physical systems* occur; this is about the fact that (say) detailed and accurate depictions of freak accidents (photographic depictions, or tape-recorded ones, or written ones, or ones stored in the physical states of human brains, or whatever) almost never *precede* those accidents *themselves*.

namical laws or not, that any such sequence whatever is *readable—against the background* or *within the context* or *relative to the framework* of the best or last or canonical metaphysical interpretation of whatever complete theory of the world is under discussion—as a *story of the physical world*); and

b. that it be *complete* (which is to say, that *all* the physical facts about the world can be read off from the full temporal set of its descriptions).

Good. Let's call whatever satisfies (a) and (b) an instantaneous physical *state* of the world.

What satisfies (a) and (b) in the *Newtonian* picture (for example) is a specification of the *positions*, at the time in question, of all the particles in the world: no specification of the positions of those particles at any *one* time, or at any *collection* of times, logically entails anything whatsoever about their positions at any *other* time; and given such specifications for *all* times, *everything* about the history of the world can straightforwardly be read off.

What typically gets referred to in the *physical literature* as an "instantaneous state" of a Newtonian-mechanical universe, of course, is a specification of the positions *and the velocities* of all the particles in the world at the time in question. But the trouble with that is just that specifications of the positions *and the velocities* of all the particles in the world at one time are *not* conceptually *independent* of specifications of the positions and velocities of all the particles in the world at all *other* times.[11] The trouble (to put it slightly differently) is that a specification of the positions and the velocities of all the particles in the world at some particular instant is *not* a specification of the physical situation of the world at that instant *alone*; it is *not* a

11. And what I mean here (and maybe this deserves to be emphasized) is *all* other times. The positions and the velocities of any set of particles at any one time are *indeed* perfectly logically and conceptually and metaphysically independent of the positions and the velocities of those particles at any particular *other* time. But suppose that I is some time-*interval* within which some particular time t happens to *fall*. Then the positions and the velocities of those particles at t will certainly *not* be logically or conceptually independent of their positions and velocities throughout the *complement* of t in I.

Think (for example) of a single particle moving uniformly to the right throughout the interval ($t = -1$ minute) to ($t = +1$ minute). And now replace $x(t)$ and $v(t)$ with $x(-t)$ and $v(-t)$, respectively. And note that this maneuver leaves x and v at $t = 0$ (but not at any other time) unchanged. And note that what this maneuver results in is a straightforward logical contradiction.

specification of the physical situation of the world at that instant *as opposed to all others*, at all!

And so the Newtonian laws of motion turn out *not* (exactly) to amount to a deterministic connection between all the states of the world at all times and any single *one* of them. What those laws amount to (if you want to be careful) is a deterministic connection between all the states of the world at all times and all the states of the world throughout any arbitrarily small time-*interval*.[12]

▲▲▲ What is it, then, for something to happen *backward*?

Simple. Suppose that the true and complete fundamental physical theory of the world is something called *T*. Then any physical process is necessarily just some infinite sequence $S_I \ldots S_F$ of instantaneous *states* of *T*. And what it is for that process to happen *backward* is just for the sequence $S_F \ldots S_I$ to occur.

▲▲▲ What is it, then, for a fundamental theory of the world to fail to distinguish between past and future?

I mentioned two ideas about that, some pages back, and talked about them as if they amounted to more or less the same thing. One was that the theory entails that whatever can happen can also happen backward, and the other was that the theory offers identical algorithms for inferring toward the future and the past. And these are actually, logically, altogether different propositions. And I want to say something about precisely what their relation is.

Both of them, of course, turn out to be true of Newtonian mechanics; and

12. The laws (that is) turn out to amount to a deterministic connection not between the positions of all the particles in the world at one time and their positions at any other time, but (rather) between the positions of all the particles in the world at one time, and the *rates* at which those positions are *changing* in the immediate vicinity of that time, and the positions at any other time.

And those rates are emphatically *not* features of the physical situation of the world at any particular *instant*. And so (on this way of looking at things) the Newtonian-mechanical laws of motion turn out emphatically *not* to be anything along the lines of a set of rules whereby the world decides, on the basis of its physical situation exactly *now*, what to do exactly *next*. But (come to think of it) they *couldn't* have been. The temporal instants (after all) form a *continuum*; there is *no such thing* as the instant *immediately after* (say) five o'clock.

it goes without saying that one can imagine *other* theories of which *neither* is true; and it turns out (more interestingly) that theories can be imagined of which one of them is true and the other isn't.

There are (to begin with) two entirely distinct ways in which a theory might *fail* to offer us identical algorithms for inferring toward the future and the past.

One—the obvious one—is for the theory in question to offer us an algorithm for calculating toward the past and an algorithm for calculating toward the future and for those two algorithms to be *different*. Here's a theory like that: somewhere in space there is a fixed blue dot. And there are particles. And the law of motion is that each of those particles invariably proceeds toward that dot as time flows forward, and that the *speed* with which any particular particle proceeds toward that dot at any particular time measured in feet per second is equal to the distance between them at that time measured in feet.

But something else can happen too.

Consider (for example) a theory like this: there are, at a *number* of points in space, fixed blue dots. And there are particles. And all the particles invariably move in perfect accord with the Newtonian laws of motion, except that at noon, on a certain particular day, as the clock strikes, each particle jumps, instantaneously, to the nearest blue dot, and thereafter proceeds onward, with its pre-jump velocity, once again in accord with the Newtonian laws, forever after.

The algorithm for determining the future positions of all the particles in the world from their present ones, and from their rates of change as time flows forward, will be perfectly deterministic on this theory. But note that this theory will yield *no algorithms whatsoever* for inferring from times *after* the noon in question to times *before* it.

And note that the two very fanciful theories we've just been talking about will both flatly *deny*—unlike Newtonian mechanics—that whatever can happen to a collection of particles can also happen backward. And as a matter of fact, it turns out that *any* deterministic theory—that deterministic theories *in general*—can allow that whatever can happen forward can also happen backward *only* if the theory offers us identical algorithms for inferring toward the future and the past, and (equivalently) it turns out that deterministic theories can *deny* that whatever can happen forward can also happen

backward only if they *fail* to offer us identical algorithms for inferring toward the future and the past.[13]

Indeterministic theories are a bit more complicated. Theories with probabilistic algorithms for inferring toward the future (that is, theories whose laws stipulate the probability that such-and-such goes on at time 2 given that such-and-such goes on at time 1) generally entail nothing about the business of inferring toward the past—and yet many such theories allow that whatever can happen forward can also happen backward.

Consider, for example, a system consisting of a single particle, which can be located in either of two boxes. And suppose that the full theory of the dynamical evolution of this system, so long as it is isolated from outside influences, is that the particle is as likely as not, over any one-second interval, to switch boxes (that is, the full theory of the free dynamical evolution of this system is that the particle's probability of now being in box 1, given that it was in box 1 one second ago, is 1/2; and the particle's probability of now being in box 2, given that it was in box 1 one second ago, is 1/2; and the particle's probability of now being in box 2, given that it was in box 2 one second ago, is 1/2; and the particle's probability of now being in box 1, given that it was in box 2 one second ago, is 1/2).

This theory will entail that the time-reverse of any physically possible free trajectory is *another* physically possible free trajectory—it will entail (as a matter of fact) that the *probability* of any physically possible free trajectory (given its initial state) will be *equal* to the probability of that trajectory's time-reverse (given *its* initial state). And yet this theory will tell us *nothing whatsoever* about the probability that (say) the particle was in box 2 one second ago given that it is currently in box 1—just as the proposition that the probability of a certain die landing on 4, given that it is a fair die, is 1/6,

13. Here's why: if whatever can happen forward can happen backward, then there is a one-to-one mapping—the mapping that takes any trajectory into its *time-inversion*—between physically possible trajectories proceeding from any given present state toward the future and physically possible trajectories proceeding from that same given present state toward the past. Thus, if there is an algorithm whereby the present state plus (say) present-to-future rates of change invariably pick out a *single* possible future trajectory, then there must be only a single possible *past* trajectory with that same present state and the equivalent present-to-*past* rates of change; and that present state, and those present-to-past rates of change, must necessarily pick that past trajectory out in accord with *precisely the same algorithm*.

entails nothing whatsoever about the probability that the die is fair given that it *does* land on 4.[14]

Of course, if any theory whatsoever offers us *both* predictive *and* retrodictive algorithms, and if those two algorithms happen to be *identical*, and if the theory in question entails that a certain process can happen forward, then it will necessarily also entail that the process can happen backward. *That's* what I'll mean, then, from here on, when I speak of a theory as being invariant under time-reversal.

▲▲▲ Good. Let's talk some (with all this now under our belts) about the candidates for a fundamental physical theory that have been taken seriously *since* Newton.

Look, for example, at the classical theory of a universe made up of electrically charged particles and electromagnetic fields. What counts as an instantaneous state of the world, according to classical electrodynamics (which is what that theory is called), is a specification of the positions of all the particles and of the magnitudes and directions of the electric and magnetic fields at every point in space. And it turns out *not* to be the case that for any sequence of such states $S_I \ldots S_F$ which is in accord with the dynamical laws of this theory, $S_F \ldots S_I$ is too. And so this theory is *not* invariant under time-reversal. Period.

And neither (it turns out) is quantum mechanics, and neither is relativistic quantum field theory, and neither is general relativity, and neither is supergravity, and neither is supersymmetric quantum string theory, and neither (for that matter) are any of the candidates for a fundamental theory that anybody has taken seriously since Newton. And everything everybody has always said to the contrary (of which more later) is wrong.

14. Perhaps this is worth spelling out in more detail. The point is just this: if some large collection of particles is known, at time *t*, to be in (say) box 1, and if it is known that none of these particles will be disturbed from the outside over the course of the next second, then it can be reliably inferred that about half of them will be in box 1, and the other half in box 2, at *t* plus one second. But if some large collection of particles is known at time *t* to be in box 1, and if it is known that none of these particles *was* disturbed from the outside over the course of the *past* second, it can *by no means* be inferred that about half of those particles were in box 1 at *t minus* one second, and the other half in box 2. Suppose, for example, that the second collection of particles is itself half of some much *larger* one, all of which were *placed*, at *t* minus one second, in box 1. Nothing about the situations of those particles at *t*, needless to say, will have any bearing whatsoever on the likelihood of *that*!

There is, though, a curious *vestige* of time-reversal invariance in all these theories. There's something about all these theories that *isn't* time-reversal invariance but nonetheless somehow *recalls* time-reversal invariance or *suggests* time-reversal invariance or *smacks* of time-reversal invariance or is capable of *masquerading*, for certain purposes, as time-reversal invariance.

Let's talk about classical electrodynamics again.

It turns out that for every sequence of instantaneous states $S_1 \ldots S_F$ which is in accord with the laws of classical electrodynamics, a sequence of the form $\tilde{S}_F \ldots \tilde{S}_1$ will necessarily be in accord with them too, where the only differences between any $\tilde{S}_K$ and its corresponding S_K have to do with where the magnetic fields are pointing. And so classical electrodynamics *does* have what you might call a *partial* time-reversal invariance, a time-reversal invariance insofar as the *positions of the particles* are concerned: classical electrodynamics *does* entail that whatever motions *particles* can execute, they can also (though under *other circumstances*, with differently directed magnetic fields around) execute backward.

And so the unbreaking of glass can be no less in accord with the laws of Maxwellian electrodynamics than the breaking of glass is, and the spontaneous heating of soup can be no less in accord with Maxwellian electrodynamics than its spontaneous cooling is, and the coming of youth can be no less in accord with Maxwellian electrodynamics than its passing is, since (when you come right down to it) what it is for glass to break or for soup to cool or for people to get older-looking is just for the particles that make them up to assume certain particular sequences of positions. And so classical electrodynamics (even though it is decidedly *not* invariant under time-reversal) is every bit as much at odds with the time-directedness of our everyday macroscopic experience as Newtonian mechanics is.

And the broad outlines of all this have remained more or less in place, or at any rate they have suffered only two further complications (of which more in a minute), ever since.

None of the fundamental physical theories that anybody has taken seriously throughout the past century and a half is (as I mentioned above) invariant under time-reversal.

Most of them *are* invariant under time-reversal, though, *insofar as the positions of particles are concerned.* For most of them (more particularly) there is some fairly straightforward transformation linking every state S_K with an-

other state $\tilde{S}_K$—a transformation which varies from theory to theory, but which in every case has the property that it leaves the positions of particles unaffected—such that if $S_I \ldots S_F$ is in accord with the theory, then $\tilde{S}_F \ldots \tilde{S}_I$ is too. And (once again) since our everyday macroscopic experience is first and foremost an experience of *the positions of material bodies*, those theories are all at odds with the time-directedness of that experience in much the same way as Newtonian mechanics is.

And there are two curious pieces of contemporary physical theory that appear *not* to be invariant under time-reversal, even in the limited sense we have just been talking about.

One concerns the decays of certain sub-atomic particles. And those decays (insofar as anybody has yet been able to imagine) have nothing whatsoever to do with the time-directedness of our everyday experience.

The other is more interesting. There is—very briefly—a problem at the foundations of quantum mechanics. And there are a variety of proposals around for modifying quantum mechanics in such a way as to make that problem go away. And it happens that *some* of those proposals (though not *all* of them, and not even *most* of them) involve violations of partial time-reversal invariance too. And *those* violations (if there turn out to *be* any, if the proposals in question turn out to be *right*) might well have something to do with the time-directedness of our everyday experience. And we will be talking a great deal about all that in the last chapter of this book.

But let's leave it aside for the time being. There might well turn out (after all) *not* to be any such violations. And if there aren't, we're going to need to figure out how to alleviate this tension, or how to *live* with this tension, or what to *make* of this tension, *without* them. And the right place to start would seem to be Newtonian mechanics, where the tension is particularly simple and stark. Whatever we manage to discover there will presumably apply in the more up-to-date cases too.

4. TIME-REVERSAL INVARIANCE IN THE PHYSICAL LITERATURE

Before we get to that, though, it ought to be acknowledged in somewhat more detail that the thrust of what was reported in the previous section is quite radically at odds with what it says in the textbooks.

To begin with, what the books say it is to specify the world's complete physical situation at a certain instant is to specify what you might call its complete *dynamical conditions* at that instant, to specify (that is) all the information about the instant in question—or all the information which can in one way or another be uniquely *attached* to the instant in question—which is required in order to bring the full predictive resources of the *dynamical laws of physics* to bear. And the trouble with that (the trouble with it—that is—*as a conception of the situation of the world at an instant*) is that dynamical conditions of the world at *different* instants can turn out, as I have repeatedly emphasized, not to be logically or conceptually or metaphysically *independent* of one another.

Take Newtonian mechanics. The dynamical conditions of a Newtonian universe at any instant are not the *positions* of all the particles in the world at that instant but the positions *and the velocities* of all the particles in the world at that instant (together, as usual, with a specification of what sorts of particles they are). And the positions and velocities of all the particles in the world at some particular instant are patently *not* logically independent of their positions and velocities at other instants; and so a *specification* of those positions and velocities at some particular instant is *not* a description of the world at that instant *alone*—it is not a description of the world at that instant *as opposed to all others*, at all![15]

Talking about *going backward* in the language of dynamical conditions can consequently be a messy business. If (for example) $D_I \ldots D_F$ is an infinite sequence of Newtonian dynamical conditions corresponding to a single free particle moving to the right, then $D_F \ldots D_I$ will correspond not to

15. Maybe it ought to be stressed here that there is nothing wrong or misleading or incoherent about Newtonian dynamical conditions *per se*, and (moreover) that such conditions can perfectly well be *uniquely attached* to particular times. Those sorts of attachments (after all) are precisely the business of differential calculus. What needs to be kept in mind is just that there is all the difference in the world between being uniquely attachable *to* some particular time and being a component of the *instantaneous physical situation of the world* at that time!

There's nothing wrong with propositions like "the velocity of particle 5 at $t = 7$ is 12 miles per hour, in the x-direction." What a proposition like that is *about*, though, is not the instantaneous situation of particle 5 at $t = 7$ *itself*, but the *rate of change* of the position of that particle in the *immediate temporal vicinity* of $t = 7$. What a proposition like that is about (to put it a bit more technically) is the limit of the rate of change of the position of particle 5 over an interval *centered* on $t = 7$, as the length of that interval goes to zero.

a particle like that moving to the *left* (which is *what it is*, after all, for a process like that to happen backward) but to nothing whatsoever, to gibberish, to a contradiction.[16] And so if what counts for you as an instantaneous physical situation of the world is (somehow) a *dynamical condition*, then turning something around (at least in certain cases) must involve something *other*, something *more*, than a mere commonsensical *inversion* of the *temporal sequence* of those situations.

The books all tell it like this: for every possible dynamical condition of the world, there is such a thing as that condition "going backward." And here we are starting to get right up to our necks in paradox. What can it possibly mean for a single instantaneous physical situation to be happening "*backward*"? Never mind. Press on. For every such condition D, there is (whatever it means) some unique condition D^* which is D's so-called time-reverse. And what it is for a process $D_I \ldots D_F$ to happen backward is *not* for the inverted sequence of dynamical conditions $D_F \ldots D_I$ to happen (which will as often as not be illogical gibberish), but for the inverted sequence of *inverted* dynamical conditions—$D_F^* \ldots D_I^*$—to happen.

And what the books have to say on the question of the precise mathematical procedure for *obtaining* D_K^* from D_K is (1) that in the case of Newtonian mechanics the procedure is "obviously" to reverse the velocities of all the particles, and to leave everything else untouched; and (2) that the question needs to be approached afresh (but with the Newtonian case always somehow in the back of one's head) in each new theory one comes across; and (3) that what it is *in all generality* for one physical situation to be the time-reverse of another is (not surprisingly!) an obscure and difficult business.

It isn't, really. If you just keep your eye on the ball (which is to say, if you're careful to represent instantaneous physical situations of the world *correctly*, if you're careful to represent them in accord with the requirements of instantaneity and completeness, if you're careful to represent them by means of the sorts of things I decided, a few pages back, to call *states*) then everything is perfectly straightforward. The way to figure out what it is for any se-

16. What $D_F \ldots D_I$ will correspond to, in this case, will be a particle whose position is constantly being displaced toward the *left*, and whose *velocity* (which is by definition nothing other than the *rate of change* of that position) is constantly pointing *to the right*.

quence of dynamical conditions $D_I \ldots D_F$ to happen backward is to translate that sequence into a sequence of instantaneous states,[17] and then write that latter sequence down in reverse order, and then translate that *inverted* sequence back into the language of dynamical conditions;[18] and whatever you end up with, when all that's done, is (by definition) $D_F^* \ldots D_I^*$. The thing is that if you've fallen under the spell of the books, if the language of states is *unavailable* to you, if you've gotten it into your head that what counts as a complete description of the physical situation of the world at a pure indivisible structureless temporal instant is (per impossible!) a *dynamical condition*, then the above analysis can never even become an object of your attention.[19]

▲▲▲ Insofar as *Newtonian mechanics* is concerned, none of this ends up causing any actual trouble. The velocities of particles, after all, are nothing but the rates of change of their positions. And so if a certain sequence of genuinely instantaneous Newtonian states $S_I \ldots S_F$ corresponds to the sequence $D_I \ldots D_F$ of Newtonian dynamical conditions, and if the prescription for obtaining D_K^* from D_K is just to turn all the velocities around, then the commonsensically backward sequence $S_F \ldots S_I$ will necessarily correspond to the backward sequence $D_F^* \ldots D_I^*$ of the textbooks. And so the

17. In all the candidates for a fundamental physical theory that anybody has taken seriously since the Renaissance, a complete history of the world's *dynamical conditions* is also a complete history of the *world*, and so a complete history of the world's dynamical conditions will correspond to exactly *one* complete history of its *states*, and so the translation we are talking about here will be completely *unique*.

But there can perfectly well (in principle) be theories in which it isn't; there can perfectly well be theories in which a given complete history of the world's dynamical conditions corresponds to *more* than one complete history of its states. In cases like that, *any one* of those latter complete histories will do.

18. That *this* translation is always unique follows from the fact that *states*, by *definition*, are *complete*.

19. Dynamical conditions aren't *necessarily* distinct from states, of course. On the two-state probabilistic theory discussed above, for example, states and dynamical conditions are identical. But on any theory which is deterministic, and which is time-reversal symmetric, and which is (in a sense that will presently be clear) *non-trivial*, they *can't* be.

To see why, think of a deterministic theory T on which the state of the world can evolve (over the course of a second, say) from S_A to S_B, and then (over the course of the *next* second) from S_B to S_C. And suppose that S_A is not the same state as S_C. And suppose that this is a theory on which whatever can happen can also happen backward. Then T must entail that there are at least *two* different states (S_C and S_A) into which the state S_B can lawfully evolve, over the course of the subsequent second. And so T must entail that *not* every state-specification is also a specification of a complete set of dynamical conditions.

textbook idea of what it is to go backward is cooked up in such a way as to amount to precisely the same thing as the commonsensical idea. And so it turns out to be a consequence of the Newtonian laws of motion, on *all* accounts, that any physical process that can happen forward can happen backward too.

But in *other* theories, and as a matter of fact in *all* the fundamental theories that anybody has taken seriously *since* Newton, the plot is a good bit thicker.

Take classical electrodynamics again. What counts as an instantaneous state of the world according to classical electrodynamics is (as I said before) a specification of the positions of all the particles and of the magnitudes and directions of the electric and magnetic fields at every point in space. And it isn't the case that for any sequence of such states $S_I \ldots S_F$ which is in accord with the dynamical laws of classical electrodynamics, $S_F \ldots S_I$ is too. And so classical electrodynamics is *not* invariant under time-reversal.

But the books tell it very differently. What the *books* count as a physical situation of the world at an instant (once again) is not an instantaneous physical state but a dynamical condition. And what counts according to classical electrodynamics as a dynamical condition is a specification of the positions and *velocities* of all the particles in the world, and the magnitudes and directions of the electric and magnetic fields at every point in space. And of course a simple inversion of any sequence of *those* which is in accord with the classical electrodynamical equations of motion gives you illogic. But there turns out to be a way of *transforming* those dynamical conditions (to wit: reverse all the velocities, and reverse all the magnetic fields, and leave everything else as it was) such that if a certain sequence of those conditions is in accord with the classical electrodynamical equations of motion, then the inverted sequence of *transformed* conditions necessarily is too. And it happens (or rather, it will come as no surprise) that the books identify precisely that transformation as the transformation of "time-reversal." And so, according to the books, classical electrodynamics is no less invariant under time-reversal than Newtonian mechanics is.

The thing is that this identification is *wrong*. Magnetic fields are *not* the sorts of things that any proper time-reversal transformation can possibly turn around. Magnetic fields are not—either logically or conceptually—the *rates of change* of anything. If $S_I \ldots S_F$ is a sequence of instantaneous states of a

classical electrodynamical world, and if the sequence of dynamical conditions corresponding to $S_I \ldots S_F$ is $D_I \ldots D_F$, and if we write the sequence of dynamical conditions corresponding to $S_F \ldots S_I$ as $D_F^* \ldots D_I^*$, then the transformation from D_K to D_K^* can involve nothing whatsoever other than reversing the *velocities of the particles*. And if *that's* the case, and if $D_I \ldots D_F$ is in accord with the classical electrodynamical laws of motion, then, in general, $D_F^* \ldots D_I^*$ will *not* be.

▲▲▲ And so (notwithstanding what all the books say) there have been dynamical distinctions between past and future written into the fundamental laws of physics for a century and a half now.

And nonetheless (and on this score the books are right), those laws are all very curiously at odds with the time-directedness of our everyday experience. And that (as I said before) is the tension I mentioned at the outset. And that's what the next couple of hundred pages will be about.

▲▲▲ CHAPTER TWO

THERMODYNAMICS

Let's pay some more attention to the time-directedness of our everyday macroscopic experience.

I mentioned three such directednesses before: a directedness of influence, a directedness of knowledge, and a directedness of ordinary physical processes like the melting of ice and the cooling of soup and the spreading of smoke and the breaking of eggs and the passing of biological youth and so on. And there happens to be a breathtakingly simple and concise and elegant and powerful characterization of that third directedness, which is called the second law of thermodynamics, and which was one of the supreme achievements of the physics of the nineteenth century, and which is what this chapter is going to be about.

▲▲▲ Note, to begin with, that the sorts of physical systems in which manifest past-future asymmetries arise *are*, invariably, *macroscopic* ones, that (more particularly) they are invariably systems consisting of *enormous numbers of particles*. Systems like that apparently have distinctive properties. And it happens that in the middle of the nineteenth century a number of investigators undertook to develop an autonomous *science* of such systems.

These guys were for the most part in the business of designing steam-engines, and so the system of paradigmatic interest for them was a box of gas.

Let's talk some about systems like that, then. Let's ask, to begin with, what *terms* are appropriate for the *description* of such a system. Let's ask what it is to give an *account* of the *physical situation* of such a system. The fullest possible such account is (needless to say) a specification of the positions and velocities and internal properties of all the particles that make up the gas and its box. From that, and from the Newtonian laws of motion, the positions

and velocities of all those particles at all *other* times can in principle be calculated. And from the full *history* of those positions and velocities everything about the history of the gas and its box can in principle be *read off*. But the calculations involved here are impossibly cumbersome. And there is patently another, simpler, more powerful, more useful, more familiar, altogether different way of talking about such systems, which is to talk about them in a language of the *macroscopic*, which is (more particularly) to talk about things like the size and shape and mass and motion of the box as a whole and the *temperature* and the *pressure* and the *volume* of the *gas*. And there is patently a possible *science* of these temperatures and pressures and volumes—a science (that is) of *macroconditions*. We *know* it to be a lawlike fact, after all, that if we raise the temperature of a box of gas high enough the box will blow up. And we know it to be a lawlike fact that if we *squeeze* a box of gas from all sides the box will get *harder* to squeeze as it gets smaller. Never mind (for the moment) that this must all in principle be deducible from Newtonian mechanics. It must be possible (or at any rate it *seems* that it must be possible, or at any rate it seemed so to these guys in the nineteenth century) to *systematize* all this *on its own;* that is, it must be possible to discover an autonomous set of so-called thermodynamic laws of such boxes of gas which directly relate volume and temperature and pressure to one another, and which make no reference whatsoever to the positions and velocities of the particles of which (as it happens) the box and the gas *consist*.[1]

1. This is worth harping on some. The situation (then) is as follows. The exact and complete description of a collection of boxes of gas at some particular temporal instant (or of *anything else* at some particular temporal instant, or of the universe *as a whole* at some particular temporal instant) is called its *microcondition*, which is what we have heretofore been referring to as its *dynamical* condition, which consists—in Newtonian mechanics—of a specification of the position and velocity of every one of the particles of which those gasses and their boxes and whatever else may happen to be around are made up. And since everything there is to say about a system like that can necessarily be *read off* of its microcondition, any *other* way of talking about a system like that, any other *language* for talking about a system like that, must necessarily amount to some sort of a *carving up* of the entirety of its set of possible microconditions into *subsets*. And there is (in particular) a carving up of that set which is characteristic of everyday human language, and there is *another* carving up of that set (a very closely *related* one, of course) which is characteristic of the discriminatory capacities of ordinary unaided human *sense organs*, and there is *another* carving up of that set (and *this*—on the face of it—is a way of coming at the business from an entirely different angle) of which there can be a simple and autonomous and robust and non-trivial *dynamical science;* and it is (to begin with) something of a miracle, it is something which is by no means guaranteed merely by there being a science of the *microconditions*, that this third sort of a carving up should have existed *at all*. And it is (you might say)

And as it turns out, there *are* laws like that.

Let's have a look at them.

▲▲▲ There is, to begin with, a thing called "heat." Things get warmer by absorbing heat, and they get cooler by relinquishing it. Heat is something that can be transferred from one body to another. When a cool body is placed next to a warm one (for example), the cool one warms up and the warm one cools down, and this is in virtue of the "flow" of heat from the warmer body to the cooler one.

But what kind of a thing *is* "heat"? It turns out (on a little reflection and a little experimentation) to be a form of *energy*.

There are any number of ways of seeing that. Look, for example, at the contraption in Figure 2.1. When the first pin is removed, the gas pushes the piston out, and the piston pushes the ball, and the ball accelerates, and (it is observed) *the temperature of the gas goes down*. In the course of the pushing, then, the ball gains *energy* and the gas loses *heat*. And it has been a very deep article of faith in physics—it has been (you might even say) part and parcel of the very *idea* of energy—that the total energy of any collection of systems that are interacting with one another is necessarily always *conserved*. And so the ball's new energy must have *come* from someplace, and the only place that immediately suggests itself is *the heat relinquished by the gas*.

Imagine, too, running the above experiment *in reverse*. The ball comes in from the right, hits the piston, pushes it in, compresses the gas, heats it up, and (at the same time) slows to a stop. The ball loses energy, and that energy must have *gone* someplace, and the only place that immediately suggests itself is the heat *acquired* by the gas.

The temperature of a gas, then (if it *has* any particular temperature—if its temperature is uniform throughout), is a measure of how much energy, of how much heat, that gas has stored up inside it.

There are two ways in which gasses are known to be able to exchange energy with their surroundings. They can exchange energy as *heat* (which is what happens, for example, when bodies at different temperatures are

---

the *essence* of thermodynamics, or the fundamental condition of the *possibility* of thermodynamics, or whatever, that these three carvings up happen to amount to more or less *the same thing*; and what that thing is *called* is the carving up of the set of possible microconditions into *macroconditions*.

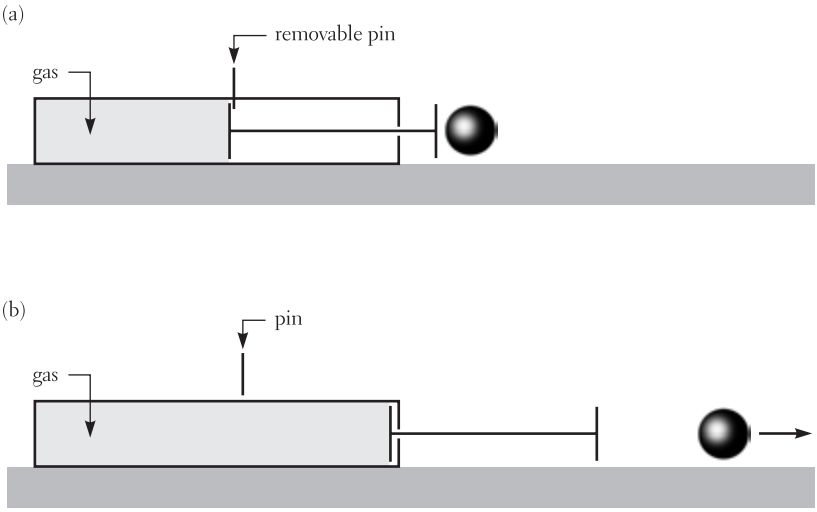


Figure 2.1

brought into thermal contact with one another), and they can exchange energy in *mechanical* form, as "work" (which is what happens when, say, the gas pushes on the ball).[2] And since energy is conserved, it must be the case that, in the course of anything that might happen to a gas,

$$DU = DQ + DW \tag{2.1}$$

where DU is the increase, in the course of the occurrence in question, of the total energy of the gas, and DQ is the energy the gas absorbs in the course of that occurrence in the form of heat, and DW is the energy the gas absorbs in the course of that occurrence in the form of mechanical work; and where DU, DQ, and DW can of course take on either positive or negative values. This (once again) is nothing other than the law of the conservation of total energy, written down in the macrolanguage of this autonomous science of

2. Maybe this deserves a little further clarification. When we speak of a gas as transferring energy to another system "as heat," we mean that the energy takes the form of heat in the system that *receives* it, and when we speak of a gas transferring energy to another system "as work," we are speaking of the gas as mechanically *pushing* on something. In both cases, of course, the heat content of the gas *itself* will go down.

temperatures and pressures and volumes; and it is referred to as *the first law of thermodynamics*.

▲▲▲ Let's sharpen our language a bit. It turns out to be convenient (to begin with) to divide the various elements of the complete macrodescription of (say) a collection of boxes of gas up into two roughly distinct *classes*, one of which contains the sorts of things that we find we are able to *arrange*, by gross everyday mechanical means, as we please (the total masses of each of the separate gasses, for example, and their total energies, and the shapes and the volumes of their boxes, and the positions of those boxes relative to one another and to other macroscopic bodies, and so on), and the other of which contains everything—or rather everything *macroscopic*—else (which is to say, a specification of the values of variables like the pressure, density, and temperature of each of the gasses as functions of position within each of their boxes). And the elements of the first class are referred to as "gross constraints" on the system, and the elements of the second—taken all together—are referred to as the system's "thermodynamic condition."[3]

Now, alterations in the gross constraints on a gas will typically bring about changes in its thermodynamic condition. If (for example) the piston in Figure 2.2 is slowly pushed in, the volume of the gas inside it will decrease, and its temperature will go up.

And it happens that if those alterations of the gross constraints are subsequently *reversed* (if, that is, the piston is slowly pulled the same distance *back out*), the volume of the gas will go back up, and its temperature will go back down, and its original thermodynamic condition will be restored. And transformations of this sort are consequently said to be *reversible*.

And the characteristic of macrosystems which will particularly interest us

3. It goes without saying that there are any number of different respects in which this sort of talk is outrageously *vague*. And the thing (for the moment) is not to *worry* about that. It turns out that—at a certain intermediate and not entirely fundamental stage of thinking things through—it is a *help*.

Take the *boxes*, for example. The boxes we've been talking about here—which is to say the *walls* of the boxes we've been talking about here—are going to have thermodynamic conditions *too*. But the usual procedure—insofar as discussions of the thermodynamics of *gasses* (such as we are engaged in now) are concerned—is to neglect those altogether. And that has turned out to be a perfectly serviceable approximation for (say) the design of steam-engines. And we will (until further notice) be adopting it here.

Figure 2.2

here is that certain of the thermodynamic transitions they undergo are *not* reversible, that certain of the thermodynamic transitions they undergo have a temporal *directionality* about them, that certain of the thermodynamic things they do when their gross constraints are altered *don't* get undone when those alterations are reversed. Consider, for example, the gas in Figure 2.3. If the wall is slid out, the volume of the gas will increase, but sliding it back in thereafter will patently have no thermodynamic effect whatsoever.

And it's with transitions like that—with *irreversible* transitions—that the *second* law of thermodynamics is concerned. What that law aims at is a very concise and very general way of summing up all we know about such transitions, a very concise and general rule for determining which transitions can be reversed, or can *occur* in reverse (in this latter way of talking about it we're taking the manipulations of the gross constraints to be a part of the process as well), and which cannot.

Let's begin to think about what such a law might look like. We can start anywhere. Let's start, then, with what is perhaps the simplest and most striking and most familiar example of an irreversible process, which is the flow of heat from a hotter body to a cooler one when the two are brought into thermal contact. The process is patently not reversible: *separating* the two bodies again will *not* cause the heat to flow *back*. Let's see, as a first shot, whether that can somehow be directly elevated to the status of a general principle.

Consider the following proposal: "heat can never flow from a cooler body to a hotter one." Note that this principle is explicitly time-reversal-asymmetric: it permits, as nature also surely does, heat to flow from hotter bodies to cooler ones. But it looks absurdly *narrow* as a candidate for a second law of the sort that we anticipated above: it seems to refer only to one of a gigantic array of *very* different sorts of irreversible processes (the spreading of smoke, the dissolving of sugar, the burning of paper, the passing of youth, and so on); it seems as if it can have nothing whatsoever to say about the *rest* of them.

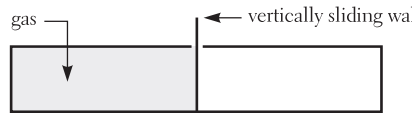


Figure 2.3

And it is going to turn out to be (as it were) the *miracle* of thermodynamics that as a matter of fact *that isn't so*.

But we're getting ahead of ourselves. The above principle, as it stands, will patently not do, because it's *false*. We know, after all, of *counterexamples:* there are such things in the world as (say) *refrigerators*. But note that it is a characteristic of refrigerators, insofar as we know, that their operation is invariably accompanied by thermodynamic changes in *the rest of the world*, that (more particularly) they require an input of *energy* in order to work, that their working requires that the total energy of the external world is going *down*. And note that the familiar phenomenon of spontaneous heat flow from hotter bodies to cooler ones requires *no* concurrent changes in the thermodynamic state of the rest of the world.

And so a principle like "no transformation whose *sole* (thermodynamic) consequence is the transfer of a given quantity of heat from a cooler body to a hotter one is possible" (which was first written down by Clausius) would seem (even if worries about the generality of its implications remain) at least to be *true*, and to point to a genuine temporal asymmetry.

Let's see what it can be parlayed into.

Consider a manifestly irreversible process which, on the face of it, has nothing to do with the sorts of heat exchanges referred to in the above-proposed second law of thermodynamics: a chair is initially sliding along a floor. There is friction. The chair slows down, and its energy of motion is converted into heat, which raises the temperature of the floor.[4] Can the Clausius law be shown to preclude the time-reverse of *that?* In the course of the *reverse* process, heat *leaves* the floor (which is to say, the floor *cools down*) and

4. This rise in temperature will of course initially be confined to those local regions of the floor which come into direct contact with the chair, but things will eventually even out, and the final, stable state of things will be one in which the chair is at rest and the floor's temperature is uniformly higher than it was when the chair was moving.

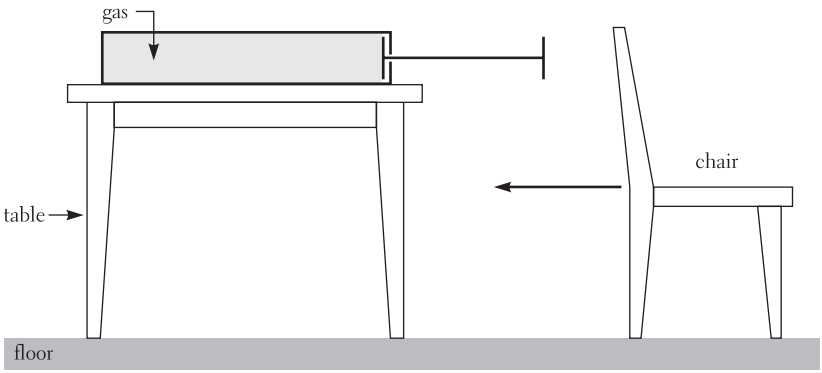


Figure 2.4

is transformed into kinetic energy of the chair (which is to say, the chair begins to *move*). And if *that* were to occur, nothing would stand in the way of our setting up a *piston* in the chair's path (as in Figure 2.4), which the chair will run into, compressing the gas inside and heating it up, while (in the process) losing its own kinetic energy. Then the gas can be irreversibly re-expanded to its original volume without further changing its temperature. And note that there is no reason at all why the initial temperature of the gas inside the piston cannot be *higher* than the initial temperature of the *floor*. And so the sole final thermodynamic result of such a process would be the transfer of a quantity of heat from one body at a lower temperature (the floor) to another body at a higher one (the gas in the piston). And so this last second-law proposal does *indeed* preclude the time-reverse of the frictional slowing down of a chair sliding across a floor.

Let's try one more. A partition is slid out (as in Figure 2.3), allowing a gas initially confined to one part of a container to expand, irreversibly, filling the entire volume. Suppose that (per impossible) that could occur in reverse. And suppose that we had prepared, in advance, a *second* box of gas, at higher temperature and at lower pressure than the first. And suppose that (at the conclusion of the hypothetical spontaneous contraction of gas number 1) we set up a two-ended piston between the two gasses as depicted in Figure 2.5. And suppose that we allow the first gas to push on that piston and compress the second (in the course of which the temperature of the first will go down and the temperature of the second will go up) until the first gas regains its

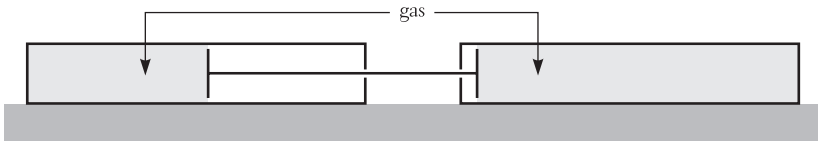


Figure 2.5

original larger volume. Finally, we allow the second gas to expand irreversibly (by the removal of another partition, say) back to its original volume.[5] Well, the net result of all this (starting with the state of things just prior to the spontaneous contraction of gas number 1) will have been to transfer heat from a cooler body (gas number 1) to a hotter one (gas number 2) with no other thermodynamic changes in the rest of the world. And so the time-reverse of the free expansion of a gas is ruled out by the Clausius formulation of the second law as well.

And so the Clausius formulation of the second law turns out to be vastly more general and more powerful than it appears at first.

▲▲▲ Let's take this a little further. The time-reverse of the chair sliding across the floor is a process in which heat is extracted from a source (the floor), which is initially at a uniform temperature throughout, and converted into mechanical energy, into *work*, and which leaves all the other thermodynamic properties of the world unchanged. And a straightforward declaration of the impossibility of anything like that turns out to be another famous formulation of the second law, the formulation due to Kelvin: "a transformation whose sole final thermodynamic result is to transform into mechanical energy heat extracted from a source which is at the same temperature throughout is impossible."[6]

5. Note, as before, that this irreversible expansion will cause no further changes in temperature, since no energy is being exchanged, here, with the rest of the world.

6. A mechanism for transforming heat extracted from a source at the same temperature throughout into mechanical energy, and which produces no other thermodynamic changes in the world in the course of that transformation, is referred to in the literature as a *perpetuum mobile of the second kind*. The *first* kind of perpetuum mobile, of course, is a mechanism for violating the conservation of energy. And so the first and second laws of thermodynamics amount, respectively, to stipulations to the effect that neither the first nor the second kind of perpetuum mobile can, in fact, exist.

Three remarks are in order here. (1) The word "sole" is just as crucial to *this* formulation of the second law as it was to the *last* one. Heat *can* be extracted from a body at the same temperature throughout and transformed into mechanical energy (for example) by putting that body in thermal contact with a gas in a container with a piston on one end, and allowing the gas to push the piston out, and allowing the piston to (say) set a ball in motion. But note that at the conclusion of this process the volume of the gas will be *larger* than it was initially, and that (of course) amounts to a thermodynamically significant difference in the state of the world. (2) This formulation, like the last one, points to a *time-asymmetry:* the transformation of *mechanical energy* into *heat*, with no other thermodynamic consequences (by means of friction between a chair and a floor, say), is a perfectly routine affair. (3) The discussion of the sliding chair shows that Kelvin's formulation of the second law is a *consequence* of Clausius's. Or rather, it shows that Kelvin's formulation is a consequence of Clausius's together with one or two auxiliary stipulations—which happen to be empirically true—to the effect that certain thermodynamic transformations (a moving chair's pushing in on a piston, for example, and compressing the gas inside, and heating it up) *are* possible.

And it turns out (given our empirical knowledge of the possibility of certain *other* transformations) that *Clausius*'s formulation of the second law is also a consequence of *Kelvin*'s.

▲▲▲ And there is yet *another* important formulation of this law, which is demonstrably equivalent to the first two, but which is (as a matter of practice) a great deal more powerful and more illuminating, and which makes reference to something called the *entropy*.[7]

Let's talk some about what that is. Consider two distinct thermodynamic states of a certain system. Call them *A* and *B*. Typically, there will be any number of different macroscopic *routes*, there will be any number of different *thermodynamic transformations*, which can get us from *A* to *B*. And some of those routes will be reversible, and some will not; and some of them may

7. That Clausius's formulation of the second law can be deduced from Kelvin's, and that *both* Clausius's and Kelvin's formulations can be deduced from the formulation we are about to discuss, and that the formulation we are about to discuss can be deduced from *either* Kelvin's or Clausius's, will all be demonstrated (along with other interesting things) in the Appendix.

involve the absorption of *heat* from the outside world, and some of them may involve the *relinquishing* of heat *to* the outside world, and some may involve the absorption of heat at *some* stages and its relinquishing at *others*, and certainly some will involve *neither*.

Good. Let Q_i represent the heat absorbed by the system during the ith stage of a certain route from A to B, and let T_i represent the temperature of the system at that stage of that route. It happens to be a consequence of both the Clausius and the Kelvin formulations of the second law of thermodynamics that the sum over all values of i of the quantity Q_i/T_i is the same for any fully reversible route from A to B as it is for any *other* fully reversible route from A to B.[8] And so the sum over all values of i of the quantity Q_i/T_i for any fully reversible route from A to B turns out to be a perfectly definite thermodynamic function of the states A and B *alone*. And the *name* of that function is the *entropy difference* between A and B. And the third and final and most powerful and most illuminating of the formulations of the second law of thermodynamics that I want to talk about in this chapter is that "the total entropy of the world (or of any isolated subsystem of the world), in the course of any possible transformation, either keeps the same value or goes up." If the transformation in question is reversible, then (needless to say) the entropy value stays constant; if the transformation is *irreversible*, the entropy goes up.

Let's think through a couple of examples.

(1) A gas, initially confined to one corner of a large container, spreads irreversibly to fill the container after a partition is removed. Note that the *temperature* of the gas (let's call it T) will be unaffected by all this, since the gas never exchanges any energy with the outside world. All right. If this new formulation of the second law is right, the entropy of the final state here had better *exceed* the entropy of the initial state. In order to see whether it *does*, what we need to do is to cook up some fully *reversible* path from the initial state to the final one. Here's one: the gas (in its initial state) is put into thermal contact with a large heat source at temperature T, and the partition is replaced by a piston, and the piston is slowly pulled out. The gas does work.

8. On many of these routes, of course, the system will absorb or relinquish heat at the same time as its temperature is continuously *changing*. Routes like that will need to be subdivided, then, into an *infinity* of distinct infinitesimal constant-temperature stages. For routes like that, the sum over i of Q_i/T_i will take the form of an *integral* over *time* of $Q(t)/T(t)$.

But (since it remains in contact with the heat source at T) its temperature remains constant. And the way it does that is by absorbing a positive amount of heat from the source. And so the entropy of the gas *rises*. And of course the entropy of the *source* will *fall*. And by precisely the same amount. And so the entropy of the entire isolated system here remains the same throughout this process. And so it must (according to this third formulation of the second law), since the transformation the system undergoes is a thoroughly reversible one.

(2) Two gasses whose masses and volumes are equal but whose temperatures are initially different (let's call them T_1 and T_2) are brought into thermal contact with each other. Their temperatures irreversibly equalize at $(T_1 + T_2)/2$. The final state had better have a higher entropy here too. Let's see. Here's a reversible path from the initial state to the final one: the gas at T_1 is put into thermal contact with a gas in a piston at temperature T_1, and the gas at T_2 is put into thermal contact with a gas in a piston at temperature T_2, and the first piston is slowly pushed in and the second piston is slowly pulled out in such a way as to equalize the temperatures at $(T_1 + T_2)/2$. Note, to begin with, that Q/T will be positive for the cooler gas here, and negative for the hotter one.[9] What about the amounts? Well, the amount of heat absorbed by the cooler gas will clearly be equal to the amount removed from the hotter one, and all the stages of the absorption occur at lower temperatures than all the stages of the removal, and so the absolute value of Q/T for the cooler gas will clearly exceed the absolute value of Q/T for the hotter one, and so the total entropy of the two-gas system will be higher at the end of this reversible process than it was at the beginning.

Note, by the way, that we have just now derived the Clausius formulation of the second law from its entropy formulation.

▲▲▲ And one more thing. Corresponding to every particular thermodynamic system, and every particular specification of gross constraints, there is exactly one stable thermodynamic *condition*, which is called the *equilibrium* condition of that system under those constraints. For example, the equilib-

9. Note that the transformations here are of the sort mentioned in footnote 8. What Q/T stands for here, then, is the sum of Q_i/T_i over the whole infinity of infinitesimal constant-temperature *stages* of these transformations; what it stands for (that is) is the *integral* over t of $Q(t)/T(t)$.

rium condition of a gas of a certain total mass and internal energy and enclosed within a container of a certain particular shape and volume is the condition in which the gas uniformly fills the container, and in which the gas's temperature and pressure are uniform throughout.[10] Any *other* condition of this gas, subject to those same gross constraints (a condition, for example, in which the gas is all concentrated in one corner of the container), will be *un*stable, and will spontaneously and irreversibly *evolve* toward the equilibrium condition, and will *stop* evolving when it gets there. And so the entropies of all the *non*-equilibrium conditions compatible with a certain set of gross constraints must necessarily be *lower* than the entropy of the equilibrium condition. And so the equilibrium condition of any particular system subject to any particular set of gross constraints will necessarily be the unique *maximal-entropy* condition of that system compatible with those gross constraints.

10. The equilibrium state of any *two*-gas system, where the gasses are in thermal contact with each other, will be one in which the temperatures of the two gasses are equal. The equilibrium state of a chair in fictional contact with a floor is one in which the chair is at rest. And so on.

STATISTICAL MECHANICS

1. THE BASIC IDEA

In the last decades of the nineteenth century, an enormously suggestive analogy was noticed between the thermodynamic properties of gasses and the *statistical* properties of large collections of *particles*, of large collections of (so-called) *molecules*.

Here's the basic idea.

Think of a gas—a gas in a box, say—as consisting of billions of tiny particles. The particles are moving around freely and more or less independently of one another. And the quantity referred to in thermodynamics as the "pressure" that this gas exerts on any particular wall of this box must presumably be a measure of the *force*—per unit time per unit area—exerted on the wall by the gas, a measure (that is) of *how many of the gas's particles* per unit time per unit area are hitting the wall, and of how *hard* they're hitting.

Good. Now consider an experiment. A gas is confined by a removable wall (as in Figure 3.1) to the left half of a large container. Its pressure is *P*. And now the wall is slid out, quickly, parallel to its surface. And the gas expands to fill the container. How should we expect this expansion to affect the gas's pressure? Well, the removal of the wall (particularly if it's removed *quickly*—so that few or none of the particles that make up the gas come into physical contact with the wall while it's in motion) will presumably not affect the *speeds* with which the particles are bouncing around. And so it will presumably not affect the *momenta* with which these particles typically *hit the wall*. But note that the average *distance* which a particle travels between collisions with a wall will *increase* when the removable wall is slid out. And so the total number of such collisions per unit time ought to be expected to

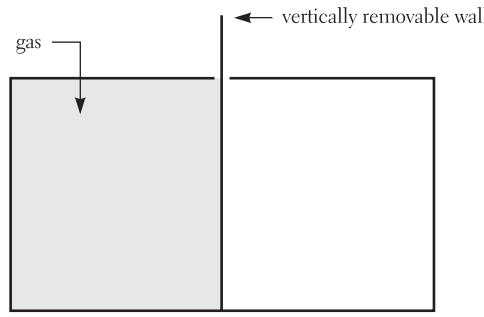


Figure 3.1

go down when the gas expands. And so the pressure ought to be expected to go down too. And so (as a matter of empirical fact) it does.

What *temperature* gets identified with, in statistical mechanics, is the average kinetic energy (that is, the average energy of *motion*, the average value of $(1/2) \times (\text{mass}) \times (\text{velocity})^2$) of the gas particles. So it turns out that heat is *not* really *another* form of energy; heat, too, is *mechanical*. *All* energy, according to the Newtonian picture of the world, is mechanical.

Consider another experiment. A gas is confined by a removable wall, just as above, to the left half of a large container. Its temperature is T. And now the wall is slid out, quickly, parallel to its surface. And the gas expands to fill the container. The sliding of the wall presumably doesn't affect the velocities of the particles, and so it presumably doesn't affect the *energies* of the particles, and so it presumably doesn't affect the heat content of the gas, and so it presumably doesn't affect the gas's *temperature*. And (as a matter of empirical fact) it doesn't.

Another. The same initial condition as above. But this time draw the wall out *slowly*, and *perpendicular* to its surface, like a piston. Now it happens to be the case, it happens to be a consequence of the Newtonian laws of motion, that a billiard ball which bounces off a receding wall (as in Figure 3.2) will move more slowly *after* the collision than *before* it.[1] And so gas particles that bounce off the wall as it's being drawn out will have their kinetic energies somewhat *depleted*. And so the temperature of this gas should go *down*,

1. Proof: consider the frame of reference in which the wall is at rest.

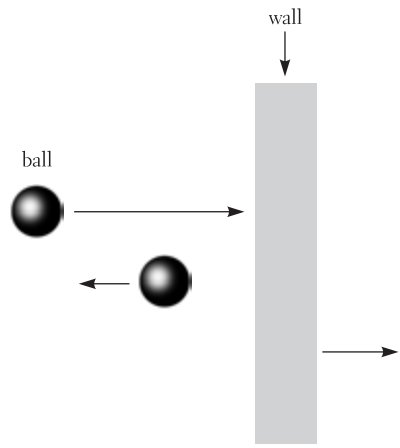


Figure 3.2

and (conversely) the temperature of a gas which a piston is pushing *in* on ought to *rise*. And so they do.

One more. Take a gas in a box. Heat it up. Keep the volume constant. The velocities of the gas particles will rise. And so the number of wall collisions per unit time per unit area will rise (and the *velocities* of those collisions will too, of course). And so the pressure ought to go up. And it does.

And there's something else. It was noticed (with microscopes) that tiny specks of dust, floating around in gasses, move in tiny, sudden, random jerks, as if they were being bombarded from all sides, as if the gas consisted of moving *particles*, of *molecules*. And it was with that (I think) that the metaphor of particles came to be taken *seriously*. It was with that that material objects came to be widely thought of as literally made up of *atoms*.

▲▲▲ What about *irreversibility*? What about the microscopic underpinnings of the *second* law?

Let's start slow.

There's an enormously illuminating thought-experiment of James Clark Maxwell's which suggests that the irreversibility of the behaviors of thermodynamic systems requires that the systems in question not be too closely *examined*, and that the irreversibility of the behaviors of thermodynamic sys-

tems must be a matter of high *probability* rather than of certainty. It's called the story of Maxwell's demon. It goes like this.

A large container is divided (as in Figure 3.3) into two separate thermally insulated chambers. One of those chambers contains a cooler gas and the other contains a warmer one. And the wall between them has a small hole in it. And the hole is covered by a small movable shutter. And the shutter is controlled by a demon (or a supercomputer or whatever—it isn't that anything *supernatural* is required here). And the demon is able to measure very quickly and very accurately the positions and velocities of all the molecules that make up the two gasses. And the way the demon runs things is this: whenever a particularly *slow*-moving molecule in the *warmer* gas approaches the shutter—a molecule whose kinetic energy is lower even than the average kinetic energy of the molecules in the *cooler* gas[2]—the demon opens the shutter and lets it *through* to the cooler gas. And whenever a particularly *fast*-moving molecule from the *cooler* gas approaches the shutter—a molecule whose kinetic energy is *higher* even than the average kinetic energy of the molecules in the *warmer* gas—the demon opens the shutter and lets it through to the warmer gas. And so the net effect of all this is to raise the average kinetic energy of the molecules in the warmer gas and to lower the average kinetic energy of the molecules in the cooler gas—to raise the *temperature* of the warmer gas and to *lower* the temperature of the *cooler one*—to transfer *heat* from the cooler gas to the warmer one. And note that all this occurs without the demon's having been required to do any *work*[3]—note (as a matter of fact) that it all occurs without any accompanying thermodynamic changes *whatsoever* in the rest of the world. And that is, needless to say, in direct violation of Clausius's formulation (and hence all the others too) of the second law of thermodynamics.

And so if any agent or automatic device were ever in fact in a position to *survey* the precise microscopic condition of the two gasses, and to *act* on that information, then the second law would be false. And so the *truth* of that law would seem to depend on there not *being* any such agents or automatic de-

2. Molecules like that will of course make up an exceedingly tiny fraction of the molecules in the warmer gas; but there will *be* some such molecules, or at any rate there are very *likely* to be some, if the number of molecules in the gas is sufficiently large.

3. He needs to move the shutter up and down, of course, but nothing on the level of principle stands in the way of the shutter's being engineered to be arbitrarily light and arbitrarily frictionless.

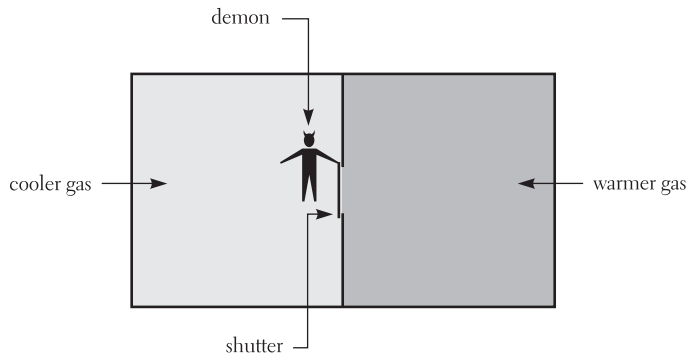


Figure 3.3

vices. And yet (and this is why the story is interesting) nothing on the level of principle would seem to *preclude* them.

Note (and this is just to belabor the obvious, but it will be useful, later on, to have written it down here) that all this depends crucially on the fact that any full specification of the *thermodynamic* situation of a gas necessarily falls very far short of being a full specification of its *physical* situation, that *thermodynamic* situations invariably correspond to enormous collections of distinct *micro*situations. It is explicitly because the demon is able to ascertain *more* than we normally can about such a gas, because he is able to ascertain *more* than is expressed by the gas's thermodynamic condition, because he is able to ascertain its *microcondition*, that he is reliably able to bring about violations of the second law.

Another story: the shudder is operated automatically, on a pre-selected schedule. It opens (say) for precisely one second precisely once every two seconds. Now clearly, it *might* be, just by *chance*, that when the shudder opens, particularly fast moving particles pass through it from the cooler chamber to the warmer one and particularly *slow*-moving particles pass through it from the warmer chamber to the cooler one. That would represent an astounding stroke of luck, of course (or at any rate, it would represent an astounding stroke of luck for it to happen *repeatedly*; it would represent an astounding stroke of luck for it to happen *to any significant degree*), but it would seem not to be altogether impossible; it would seem not to be altogether out of the question.

The idea (presumably) is something like this: the microcondition of this pair of gasses must be one or another of the enormous collection that is compatible with its *thermodynamic* condition, with its *macrocondition*; but of course we can have no idea (given only the thermodynamic information) *which* one. And there are only relatively few such microconditions which will result, on the above shutter schedule, in a net transfer of heat from the cooler chamber to the warmer one; and there are many many more that won't. And that would seem to suggest (but we will want to be thinking this particular move through a great deal more carefully later on) that the probability of such a transfer, under such circumstances, although it isn't zero, is *small*.

▲▲▲ All right. Before we get in any deeper, there are one or two further things we will need to know about Newtonian mechanics.

Consider the family of paths in three-dimensional space which a single particle, moving in some particular external field of force, moving in accord with the Newtonian laws of motion, might traverse between $t = -\infty$ and $t = +\infty$. That family will consist of some infinite collection of continuous curves, going every which way, crossing over one another, and perhaps single *points* as well (which represent cases in which the particle is permanently at rest). Consider, for example, the family of paths which a single *free* particle, a single particle subject (that is) to *no* external forces, might traverse. *That* family will consist of every single straight line *there is* in three-dimensional space, and every single *point* there is too.

And all this is a bit of a mess. And it happens that there is a much prettier and more informative way of representing things. In order to specify the position of a particle in three-dimensional space, we need to specify three numbers; we need to specify the values of the particle's three *coordinates*. And in order to specify the *velocity* of a particle in three-dimensional space we need to specify three *other* numbers (the three *speeds* with which the particle is progressing along the x, y, and z axes, respectively). And that suggests a way of representing the full dynamical conditions of a single-particle system, in a three-dimensional space, at an instant, as follows: think (instead) of a *six*-dimensional space, and represent the full dynamical conditions of a single-particle system at an instant by a *point* in that space, using the first three of its coordinates as *position* coordinates and the *second* three of its coordinates as

velocity coordinates. A space like that is referred to in the literature as the *phase space* (as opposed to the three-dimensional *position space*) of the single-particle system in question.

Picking out a point in phase space, then, corresponds to a full specification of the dynamical conditions of a single-particle system at an instant, and we can of course plot out possible *trajectories* (that is, we can plot out possible continuous *sequences* of dynamical conditions; we can plot out possible continuous sequences of positions and velocities) in phase space too.[4]

Consider how those trajectories will look. Here are a few observations: think (to begin with) of a point in ordinary *three*-dimensional space, in *position* space, at which two possible position-space trajectories of a single-particle system *cross*. Think (that is) of a point in position space which two possible position-space trajectories *share*. The *velocities* associated with those two trajectories at that common point must clearly be *different*, since the two trajectories proceed to different points in position space slightly later on. Indeed, we know that the position and the velocity of a single-particle system at any one instant completely determine that system's entire future and past, which is to say that they completely determine that system's entire *trajectory*, and so no two trajectories that are in any respect distinct can *ever* share both a position and a velocity.

The determinism of the Newtonian laws of the motions of single-particle systems, then, entails that no two trajectories, depicted in *phase* space, can ever *cross!* Trajectories depicted in phase space will tend to flow along side by side, as in Figure 3.4. The family of phase-space trajectories of a *free* single-particle system, which moves about (for simplicity) in a *one*-dimensional position space, is depicted in Figure 3.5. Note that each of the individual points on the *x*-axis is a full trajectory in and of itself.

The generalization to many-particle systems is simple. One point in $6N$-dimensional space can represent all the positions and velocities (which is to say, it can represent the complete dynamical conditions) of a system of N particles moving around in a three-dimensional position space. The motion of that single point in the phase space describes, in every detail, the individual motions of all N particles, including a specification of *which particular particles* are moving which way. And as before, the determinism of the

4. Every particle (after all) at every instant has some position and some velocity!

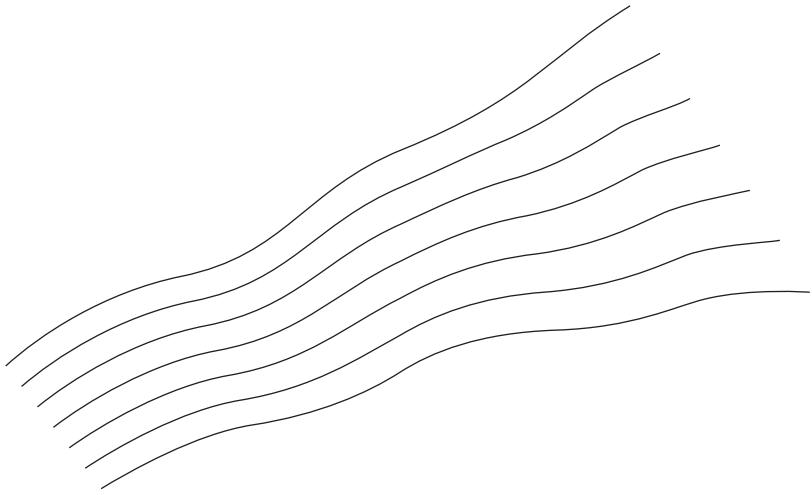


Figure 3.4

Newtonian laws of the evolutions of such systems will entail that no two paths in the phase space of any isolated dynamical system ever cross.

And yet another representation will come in handy sometimes. *N* points, *N numbered* points, in a *six*-dimensional space (rather than a 6*N*-dimensional one) can pick out the complete dynamical conditions of an *N*-particle system as well. Their *N* trajectories will specify the motions of all the particles. Here, the implication of classical determinism will be that no two distinct *complete sets of N trajectories* can ever *all* coincide at any single particular moment. Spaces like that are called *mu*-spaces. For single-particle systems, of course, phase space and mu-space coincide.

▲▲▲ Good. Let's get back to statistical mechanics.

The various different possible *macroconditions*, the various different possible *thermodynamic* conditions of any particular physical system, will presumably correspond to different *regions* of that system's *phase* space. Everyday macroscopic human language (that is) carves the phase space of the universe up into *chunks*.

And the heart of the statistical-mechanical account of the second law of thermodynamics is the observation (which was originally Boltzmann's) that

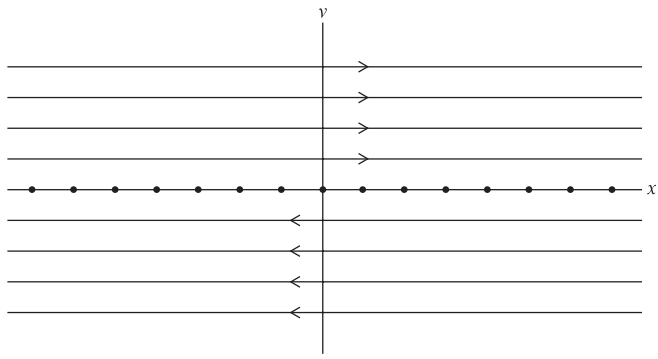


Figure 3.5

this carving up is radically *uneven*, that some of the chunks are radically *larger* than others.

Think (to begin with) in *mu*-space. Think (say) of the mu-space of a gas in a rectangular container. And divide that space up into equal-sized cubical *boxes*. And call a specification of *which* of the N identical gas particles is in which of those equal cubical boxes an *arrangement*. And call a specification of *how many* particles there are in every particular box (but not of *which* particular particles are in every particular box) a *distribution*. And so a distribution will typically be compatible with a number of *different* arrangements. And so a distribution will typically convey much less *information* than will an arrangement. And clearly every distinct *arrangement* of this gas will be compatible with any one of an *infinite* collection of its *microconditions*. Every distinct arrangement of this gas (that is) will be compatible with any one of the infinity of points in some finite *region* of its *phase* space. And a little reflection (of the sort that goes on in Figure 3.6, for example) will show that the *volume* of the region corresponding to any *one* arrangement will be equal to the volume of the *different* region—the *disjoint* region—corresponding to any *other* arrangement.

And note (finally) that if the dimensions of the boxes are microscopic, and if the boxes are nonetheless large enough so that there are many fewer of them than there are particles in the gas, then what we've been referring to as a *distribution* and what we've been referring to as a *macrocondition* will come to more or less the same thing: they will both (that is) amount to specifications of pressure and temperature and density and momentum and

Consider (for simplicity) a system consisting of two particles, each of which is free to move about only in one spatial dimension, and both of which are (moreover) confined within a certain finite *interval* of that dimension. And suppose that what it means to give an *arrangement* of these two particles is simply to specify about each one of them separately whether it happens to lie in the right half or the left half of that interval. And forget altogether (just for the moment, just so as to make everything representable on a two-dimensional sheet of paper) about the *velocities* of these two particles. Then the correspondence between arrangements and compatible regions of phase space is going to go like this:

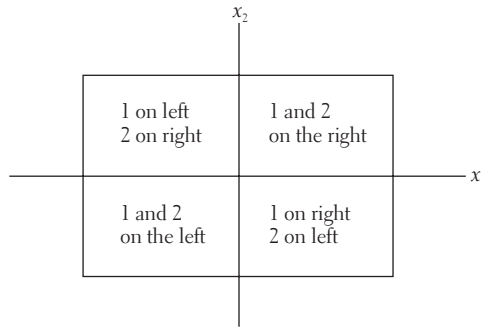


Figure 3.6

energy and charge and chemical composition and what have you as functions of approximate location, as functions (if you will) of *coarse-grained* location, in ordinary geometrical space.[5]

Good. Now the punch line. What Boltzmann observed was this: the distribution in which all N of the particles are located in box number one (or in box number two or in box number three or in *any* particular box) corresponds to exactly *one* arrangement. And the distribution in which $N - 1$ of the particles are located in box number one and one of the particles is located in box number two corresponds to N arrangements. And the distribution in which $N - 2$ of the particles are in box number one and two of the

5. The boxes need to be microscopic, then, so as to be capable of standing in, for all *macroscopic* intents and purposes, as "points"—and as to the boxes nonetheless being large enough so that there are many fewer of them than there are particles in the gas, the idea *there* is that an individual box should typically contain either a statistically significant number of particles or none at all, so that talk of statistical properties like the "temperature" or the "pressure" or the "density" of the gas at a particular coordinate space box makes good sense—and so it turns out to be part and parcel of what thermodynamic systems *are* that they are the sorts of systems whose mu-spaces can be divided up into equal-sized boxes which simultaneously satisfy the above two constraints.

particles are located in box number two corresponds to $N^2 - N$ arrangements. And the distribution in which $N - 3$ of the particles are located in box number one and three are located in box number two corresponds to $N(N - 1)(N - 2)$ arrangements. And (more generally) *dispersed* distributions correspond to larger numbers of arrangements (and *vastly* larger numbers, mind you, if N is large) than *concentrated* distributions do.

And recall that we set things up so that the volume of the chunk of phase space corresponding to any particular arrangement is equal to the volume of the chunk of phase space corresponding to any *other* particular arrangement. And so the upshot of all this is that *dispersed distributions correspond to vastly larger chunks of phase space than concentrated ones do.*

And it turns out to be precisely this imbalance (as it were) that gets the statistical-mechanical account of the second law of thermodynamics off the ground.

▲▲▲ But before that story gets told, a little digression is in order, on the question of precisely where this imbalance *comes* from.[6] The argument above (which is the canonical one and the simplest one and the one that's in all the textbooks and the one that everybody learns as an undergraduate) is dangerously misleading about that. It makes it appear as if the imbalance depends on there being some determinate matter of fact about *which particular particle* is in *which particular mu-space location;* it makes it appear as if the imbalance depends (say) on the two situations depicted in Figure 3.7 being physically or metaphysically or in some meaningful way actually *distinct* from each other. It makes it appear (that is) that the imbalance depends on an explicitly *Haecceisstistic* calculation of volumes in phase space.[7] And as a matter of fact, it *doesn't* depend on that. And that deserves to be made very clear.

Let's start slow. The above calculation of volumes was carried out (once again) in an explicitly Haecceisstistic phase space, the sort of space in which the situations depicted in Figure 3.7 correspond to *two distinct points.* And

6. And what follows here, by the way, is due to Nick Huggett, who published a very elegant paper, "Atomic Metaphysics," about all this a year or so ago in the *Journal of Philosophy* (96, no. 1 [January 1999]: 5–24).

7. Haecceisstism, then, is the doctrine that two worlds which differ from each other by means of nothing over and above a simple permutation of the positions of otherwise identical material particles are (nonetheless) *different.*

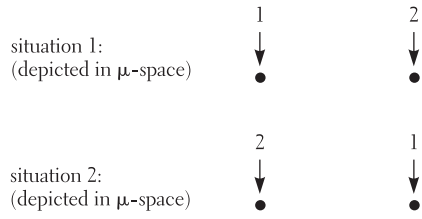


Figure 3.7

what immediately jumps out at everybody who looks at that calculation is that switching to a *non*-Haecceisstistic phase space, switching (that is) to a space in which the situations depicted in Figure 3.7 correspond to just a *single* point, will drastically *reduce* the phase-space volumes associated with dispersed macroconditions. And so it will. But what's easier to *miss* (but no less *true*) is that switching to a non-Haecceisstistic phase space will reduce the volumes associated with *non*-dispersed macroconditions *as well*, and by *precisely* the same factor.

Let's see how that works. Consider a system consisting of two particles. And suppose that each of those particles is free to move about in only a single spatial dimension. And forget (for the moment) about their velocities. The spatial configuration of a system like that can be represented, Haecceisstistically, by a point in a two-dimensional phase space like the one depicted in Figure 3.8, which is divided (as above) into separate boxes. The points in box α correspond to situations in which particle one is in region *A* and particle 2 is in region *B*, and the points in box β correspond to situations in which particle one is in region *B* and particle two is in region *A*, and the points in box γ correspond to situations in which both particles are in region *A*, and the points in box δ correspond to situations in which both particles are in region *B*. And a very natural *non*-Haecceisstistic phase space for this system can be carved out of this Haecceisstistic one simply (say) by discarding the region below the diagonal dotted line.

Of course, switching to a non-Haecceisstistic method of counting up the possible configurations of a pair of particles like these will simply *do away* with the distinction between what we've been calling an *arrangement* and what we've been calling a *distribution*. Switching to a non-Haecceisstistic method of counting up the possible configurations of a pair of particles like

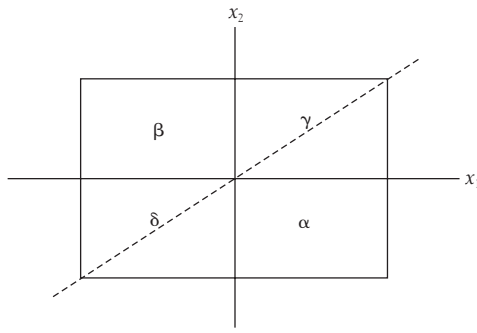


Figure 3.8

these (that is) will make it the case that there are exactly as many arrangements of these particles in which one of them is in *A* and the other is in *B* (namely: 1) as there are arrangements in which (say) both of them are in *B*. That's the thing that jumps out at everybody right away. What seems to have been easier to miss (but which is right there to see, for example, in Figure 3.8) is that switching to a non-Haecceisstistic counting method will *also* do away with the principle that different arrangements correspond to *equal volumes of phase space*. Notwithstanding (for example) that there are exactly as many non-Haecceisstistic arrangements of these particles in which one of them is in *A* and the other is in *B* as there are non-Haecceisstistic arrangements in which both of them are in *B*, the *volume of the region of non-Haecceisstistic phase space* in which one of these particles is in *A* and the other is in *B* is *twice as large* as the volume of the region in which both of them are in *B*, *just as it was in the Haecceisstistic case*. And so it will go in general. And so (notwithstanding everybody's impression to the contrary) the imbalance we have been talking about, the imbalance which is crucial to the statistical-mechanical account of the second law of thermodynamics, has nothing whatsoever to do with the question of Haecceisstism.[8]

8. There's a certain fairly trivial sense in which it ought to have been obvious from the outset (if we had stopped to think about it) that the facts of thermodynamics cannot possibly shed any light on the truth or falsehood of the doctrine of *Haecceisstism*. The question of the truth or falsehood of the second law of thermodynamics is (after all) a straightforwardly *empirical* one; and the question of Haecceisstism, the question (that is) of whether or not certain *observationally* identical situations are identical *simpliciter*, manifestly is *not*.

Nonetheless, it might have turned out that the statistical-mechanical account of thermody-

That having been said, we will, for the most part, be working with explicitly Haecceisstistic phase spaces here. They are (if only for the purposes of smoothly connecting up with the physical literature) a good deal more convenient. But it will need to be remembered that that in no way, shape, or form represents a matter of principle.

▲▲▲ All right. Let's get back to our story. Let's figure out what to make of Boltzmann's observation.

Consider, to begin with, how the *entropy* of any given macrocondition depends on the shape of its *distribution*.

Consider, for example, two macroconditions of a gas, in one of which the gas is concentrated in one corner of a large container, and in the other of which it is more or less uniformly dispersed *throughout* the container, and in both of which the temperature of the gas is the same. We learned in Chapter 2 that the entropy of the dispersed condition is higher than the entropy of the concentrated one. And note that the condition that's more dispersed in ordinary *coordinate* space (since the temperatures of these two conditions are the same, which is to say that the average *kinetic energies of the gas particles* in these two conditions are the same, which is to say that the *momentum* distributions associated with these two conditions are the same) will be more dispersed in *mu*-space as well. And so (in this case, at least) the higher-entropy macrocondition is the one that corresponds to the larger number of arrangements, and the higher-entropy macrocondition is the one that corresponds to the larger volume in phase space.

This is suggestive. Let's look further. Consider two macroconditions of a gas, in both of which the gas is more or less uniformly dispersed throughout its container, but in one of which its temperature is high and in the other of which its temperature is low. The higher-temperature condition will be the higher-entropy one. And the coordinate space distributions of these two conditions will (by stipulation) be the same. And of course the momentum distribution of the warmer condition will be more dispersed than the momentum distribution of the cooler one. And so the overall mu-space distribution

namics is somehow radically *simpler* or more *natural* or more *compelling* or more of an *explanatory success* when expressed in a *Haecceisstistic* language than it is when expressed in a *non*-Haecceisstistic one. And the thing we've just learned (which seems to me substantive and non-trivial and impossible to have anticipated without doing the work) is that that is not the case.

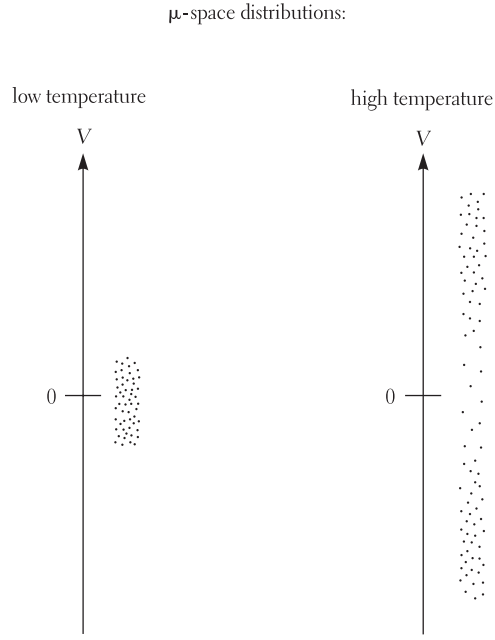


Figure 3.9

associated with the higher-entropy condition will once again be more dispersed (as depicted in Figure 3.9) than the overall mu-space distribution associated with the *lower*-entropy one. And so the higher-entropy macrocondition will again be the one that corresponds to the larger number of arrangements, the one that corresponds to the larger volume in phase space.

Let's try one more. Consider two macroconditions of a *pair* of gasses, each of which is uniformly dispersed throughout its container. In one condition the two gasses have different temperatures, and in the other their temperatures are equal. The latter condition is (as we learned in the last chapter) the higher-entropy one, and it turns out here too (although the demonstration is a bit more involved) that the latter condition is the one that corresponds to the larger number of arrangements and to the larger volume in phase space.

And so on. And this is Boltzmann's first great insight: that for *all* thermodynamic systems, looked at in an appropriately coarse-grained mu-space,[9]

9. In the sense described in footnote 5.

higher entropies correspond to larger numbers of arrangements, and larger volumes of phase space.

▲▲▲ And as a matter of fact, Boltzmann was able to find a stunningly simple and general mathematical expression for the thermodynamic entropy of a distribution as an explicit function of the number of arrangements compatible with it, which is:

$$S = K \log n \tag{3.1}$$

where S is the entropy of the distribution in question, n is the number of arrangements compatible with that distribution, and K is a number known as Boltzmann's constant.[10]

And so we now have in hand a variety of new and beautiful and profound and statistical ways of thinking about entropy.

To begin with, the entropy of a macrocondition is a measure of something like the *number of microconditions* compatible with a given macrocondition. That can't be *exactly* right, of course: the possible microconditions of any system (after all) form a continuum, and every macrocondition will necessarily be compatible with an infinity of them. What the entropy measures, then, is not quite the number of *microconditions* compatible with the macrocondition in question, but (as we saw above) the number of *arrangements* compatible with it, the *volume* of the infinity of compatible microconditions in *phase space*.

And of course the fact that the entropy of a macrocondition is a measure of the volume of the infinity of microconditions compatible with it, the fact that the entropy of a macrocondition is in some sense a measure of *how*

10. The precise mathematical form of this function is (of course) determined by the requirement that it match the thermodynamic entropy for all thermodynamically well-defined circumstances.

Consider, for example, why the function needs to be proportional to the *logarithm* of n rather than to n *itself*. It's like this. The thermodynamic entropy of any *collection* of thermodynamic systems is (think about it) the *sum* of the entropies of each of those systems separately. But the number of *arrangements* compatible with the macrocondition of any such collection will be the *product* of the numbers of arrangements compatible with the macroconditions of each of those systems separately. And so it will be the *logarithms* of the numbers of arrangements compatible with the macroconditions of each of those systems, and not the numbers of such arrangements *themselves*, that add up the way thermodynamic entropies do.

many microconditions are compatible with it, means that entropy has something to do with *information*. Entropy (that is) is a measure of *how much one can infer* about a system's microcondition from knowledge of its macrocondition. The higher the entropy of a macrocondition, the larger the volume of phase space which is compatible with it, the larger the number of microconditions which are compatible with it, the less information that macrocondition carries, the less a knowledge of that macrocondition can *tell* you.

And if (say) all we know for certain of some particular system at some particular instant are its gross constraints, and if (for whatever reason, of which more later) the probability we assign to the system's being in any particular one of the microconditions compatible with those constraints is equal to the probability we assign to the system's being in any *other* particular one of the microconditions compatible with those constraints,[11] then entropy is patently a measure of *probability*, then (more particularly) higher-entropy macroconditions will be *more probable*, will be *much* more probable, than lower-entropy ones.[12]

And entropy clearly has something to do with (at least) intuitive ideas of *randomness* and *disorder*. Conditions with higher entropies are in some sense less *structured*, less *arranged*, less *bunched up*, more *dispersed*, more of a *mess* than those with lower entropies.

▲▲▲ Now we're getting somewhere. None of this (as yet) is explicitly *dynamical*, but dynamical implications are patently not all that far off.

Let's think some about how to get *at* them.

Here (to start with) are three crude stabs:

(1)

Suppose that a certain system is at present in a certain non-maximal entropy macrocondition. Suppose (for example) that a gas is concentrated, at present, in one corner of a large container. And suppose we would like to get an idea of how this gas is going to be distributed in the coordinate space dimen-

11. That is, if (for whatever reason, of which more later) our probability-distribution is *uniform* over the entire region of the phase space of the system which is compatible with the gross constraints.

12. Since entropy is proportional to the *log* of the volume—and hence also to the log of the *probability*—of the macrocondition in question.

sions of its mu-space one second from now. And suppose we would like to get this idea without going to the trouble of finding out exactly what the present microcondition of the gas *is*, and without going to the trouble of actually applying the Newtonian equations of motion to a system consisting (as this one does) of a huge number of particles.

Well, here's something we might do—it's crude (as I warned), but it has a compelling sort of reasonableness about it: consider (as Boltzmann did) the full set of coordinate-space distributions of the gas particles which might *possibly* obtain a second from now given the gas's initial macrocondition (given, more particularly, the average initial *speed* of the particles of which the gas consists). There are a number of such distributions (see Figure 3.10). But note that one of those distributions (the maximally uniform, maximally dispersed one) is associated with an overwhelmingly larger volume of *phase* space, with (as it were) an overwhelmingly larger number of micro*destinations*, than any of the others.[13] And since we have no idea whatsoever *which particular one* of the microconditions compatible with any of the above-mentioned distributions is the one that this gas will actually *assume*, it would seem to make sense to count every one of those microdestinations as, a priori, *equally probable*. And of course *that* will mean that it is overwhelmingly likely that the distribution that this gas assumes one second from now is precisely the maximally uniform maximally dispersed one pictured above—precisely the distribution which our experience (and the summation of that experience in thermodynamics) informs us it *does* assume!

And the same reasoning will now entail that the uniform spreading is very much to be expected to *continue* for the *subsequent* second, and thereafter as well, until the gas finally uniformly fills the entire container. And at *that* point, what the same sort of reasoning will entail is that the container is overwhelmingly likely to *stay* uniformly filled.

And so we have apparently succeeded here in deducing (from what seem like entirely innocent and reasonable assumptions about the behaviors of microsystems) an *irreversibility*! We have succeeded in deducing (that is) that whereas concentrated distributions of this gas can be expected to evolve

13. And note that this is entirely independent of the distribution of the gas particles in the *velocity* dimensions of the mu-space. For *every* such distribution in velocity space, the coordinate space distribution that takes up by far the largest volume of *phase* space will be the one that's maximally uniform and maximally dispersed.

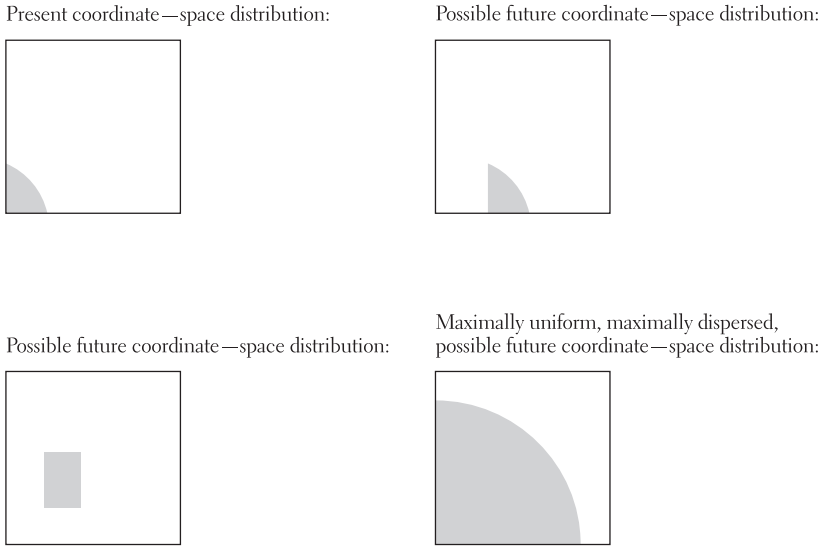


Figure 3.10

into dispersed ones, dispersed ones are *not* to be expected to evolve into concentrated ones.

(2)

The gross constraint that typically comes into play in the position dimensions of the mu-space is (as we've seen) something to the effect that each of the particles in the gas is located within some specified *spatial region*, within some particular *container*—a constraint (that is) on the values of certain physical variables of each of the particles *separately*.

The *velocity* dimensions, by contrast, tend to be a bit more complicated. The constraints you encounter *there* are generally something along the lines of a fixed average energy per gas particle[14]—something (that is) that will give rise both to constraints on each of the particles separately and *also* to mathematical relations *among* them.

And it turns out that the most convenient way of thinking through the consequences of relations like that is to think them through in the limit at which the number of particles in the gas goes to infinity, and the sizes of the boxes

14. That is, a fixed total energy for the gas as a whole.

in momentum space go to zero, so that the specification of a distribution amounts to the specification of a continuously defined *density-function*.[15] And the density-function which turns out to maximize the entropy, in the limit as N goes to infinity, with a fixed average per-particle energy (that is, the density-function which turns out to occupy the largest volume in the phase space, the density-function which has the least internal structure, the density-function which is the most dispersed, in the limit as N goes to infinity, with a fixed average per-particle energy) is the so-called *Maxwell-Boltzmann* distribution, which is a bell-shaped curve that *peaks* at that average.

Think (then) of a gas which is composed of an infinite collection of particles. And consider the rates at which those particles will typically be altering the magnitudes and directions of one another's *velocities*—altering them (that is) by means of *collisions*. Consider (more particularly) the rate at which pairs of particles in that gas, at some particular location x in physical space, can be expected to undergo the particular sorts of collisions in which (as depicted in Figure 3.11) the velocity of one of them gets changed from v_1 to v_1' and the velocity of the other one gets changed from v_2 to v_2'.

To begin with, that rate (as a little reflection will show) must always be expressible in the form $f_x(v_1)f_x(v_2)[v_1 - v_2]C$, where C (which is known as the *cross-section* for collisions of the type $(v_1,v_2 \rightarrow v_1' ,v_2')$) is deducible from information about the structure of the *interactions* between the particles in question.[16]

And suppose (and this is just to suppose that the interactions between the particles in question are *invariant under rotations*) that the cross-section for collisions of the type $(v_1,v_2 \rightarrow v_1' ,v_2')$ happens to be identical to the cross-section for collisions of the type $(v_1' ,v_2' \rightarrow v_1,v_2)$.

15. Of course, the right way of expressing the energy constraint in the limit as the number of particles in the gas goes to infinity will be in terms not of the *total* energy of the gas, but of its per-particle *average*.

16. Here's the idea: to begin with, $f_x(v_1)$ and $f_x(v_2)$ are the initial densities of particles at the (x, v_1) and (x, v_2) coordinates (respectively) in the mu-space. Patently, the rates of the sorts of collisions we're talking about must go up, linearly, as either of those go up.

Moreover, $v_1 - v_2$ will be the rate at which particles of the sorts we're talking about will initially be *approaching* one another, and the rates of the sorts of collisions we're talking about will clearly be proportional to that too.

And then there will be a factor, C, which depends on precisely how the particles in question *interact* with one another, and which will determine the frequency with which such interactions deflect particles from (v_1, v_2) to (v_1', v_2').

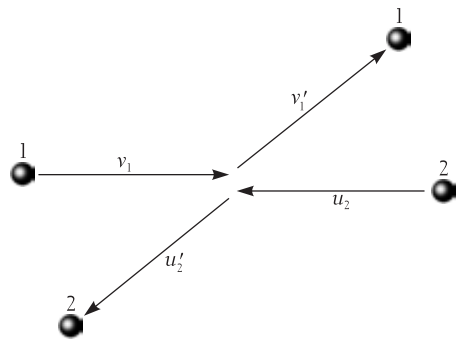


Figure 3.11

And suppose (and here we are dealing with something a bit less innocent, a bit less obviously true) that the values of the density functions $f_x(v)$ happen *not to depend on x*; suppose (that is) that at the moment in question there happen to be no systematic correlations whatsoever between the velocities of particles in this gas and their locations in ordinary physical space.

What Boltzmann was able to prove rigorously is that if all the above suppositions are true, then the velocity space density-function of this gas necessarily either *is* the Maxwell-Boltzmann density-function at the moment in question or is evolving *toward* the Maxwell-Boltzmann density-function, at that moment, as time flows forward. That's his famous "H-theorem." And so (given all we've supposed, of which there will be a good deal more to say later on) we have irreversibility, and we have entropy increase, and we have the approach to equilibrium, all over again.[17]

(3)

Let's take another tack. This one (which is more in the tradition of the American physicist J. W. Gibbs) is best laid out in the *phase* space.

In the phase representation (remember) the full Newtonian dynamical conditions of an *N*-particle system at any particular instant correspond to a *single point* in a 6*N*-dimensional space. And if the system in question is

17. The upshot of all this, then, is that collections of particles will tend to *spread out*, as much as the gross constraints allow, in mu-space. And the *reason* is always (roughly) that the maximally spread-out distributions correspond to by far the largest number of microdestinations.

known to have some particular *total energy*, that (of course) will *restrict* the possible locations of the system-point in the phase space.

Consider, for example, a system consisting of a single free particle, which moves about only in a single spatial dimension. The phase space of a system like that will be *two*-dimensional, and the information that the total energy of that system is (say) E will restrict the possible locations of the system-point to one of the two lines in Figure 3.12a.[18]

Or consider a system consisting of *two* free particles, each of which moves about only in a single spatial dimension. The phase space of a system like *that* will be *four*-dimensional. And the information that the total energy of that system is E will here restrict the possible locations of the system-point to the four-dimensional counterpart of a *cylinder*; the information that the total energy of that system is E (that is) will restrict the possible locations of the system-point to a region whose cross-section at every possible combination of the position values of the two particles is precisely the *circle* in Figure 3.12b.

And that's how things *always* go: the set of locations in any Newtonian system's phase space which is compatible with that system's having any particular total energy E forms some continuous *region*, and the *dimensionality* of that region is typically larger than two, and the dimensionality of that region is invariably *less by one* than the dimensionality of the phase space itself, and that region is consequently referred to in the literature as the system's "energy-E hypersurface." And the conservation of the total energies of isolated Newtonian systems clearly entails that the trajectory of any such system can never wander off the particular energy hypersurface it *starts out* on.

Good. Let's get back to gasses.

Consider (then) a gas which is subject to some particular set (*any* particular set) of *gross constraints*—constraints like the *size* of the box that the gas is in, and the *location* of that box, and the total *energy* of the gas particles, and so on. We know from our mu-space discussions that the overwhelming majority of the microconditions of such a gas are conditions whose entropy value is precisely at the *top* of the range that those constraints allow. We

18. The $6N$-dimensionality of the phase space that I referred to above presumes that each of the N particles is free to move around *three* ordinary spatial dimensions. The more general rule is that the dimensionality of the phase space of an N-particle system is $2 \times N \times$ (the number of ordinary spatial dimensions in which the particles in the system are free to move around).

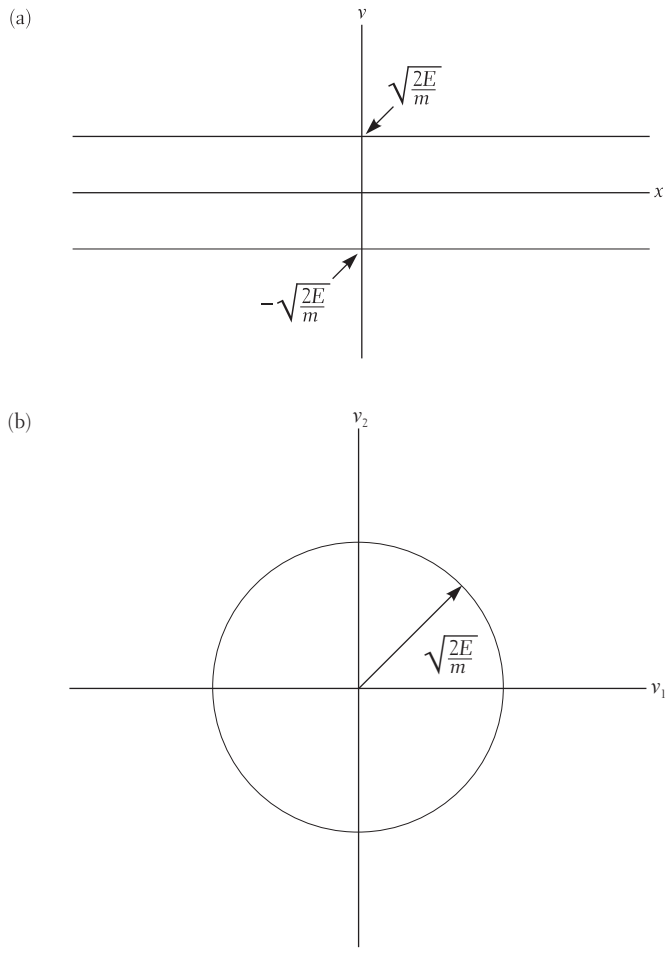


Figure 3.12

know (that is) that the overwhelming majority of the microconditions of such a gas are conditions corresponding to *equilibrium*.

And what that amounts to in *phase* space (which is depicted in Figure 3.13) is that equilibrium conditions take up the overwhelming majority of the *surface area* of the relevant *energy hypersurface* of a gas like that.

Consider, then, a gas which is known to be concentrated at a certain initial time in one corner of a large container. Consider (that is) a gas whose sys-

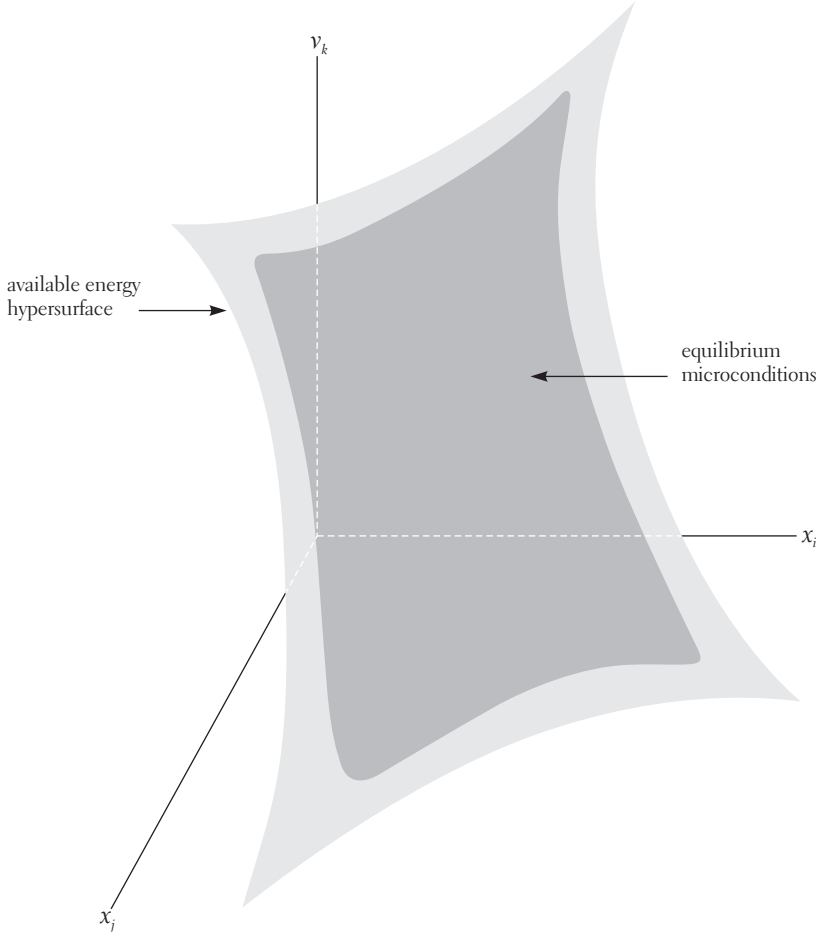


Figure 3.13

tem-point is known to be located at a certain initial time in a certain particular *non-equilibrium* region of the sort of hypersurface depicted in Figure 3.13. And suppose (once again) that we would like to get some idea of how that gas is going to evolve; suppose that we would like to get some idea of where that system-point is going to *go*, over the next ten minutes or so. And suppose (as before) that we would like to get this idea without going to the trouble of finding out exactly what the present microcondition of te gas *is*, and without going to the trouble of actually applying the Newtonian equa-

tions of motion to a system consisting (as this one does) of a huge number of particles.

Well, something we might do, something that seems to have a compelling *reasonableness* about it (more or less in the spirit of what we were doing on page 52), is to suppose that the system-point wanders aimlessly, randomly, every which way, in no particular direction, favoring no particular region, all over the energy hypersurface. Let's see what *that* can be parlayed into.

Let's start by sharpening it up some. Let's say of a system-point's trajectory that *it favors no particular region of its available energy hypersurface* if, for all choices of T, in the limit as the length of a time-interval centered on T goes to infinity, the fraction of that interval that that trajectory spends in any particular region of that hypersurface is equal to the area of that region divided by the area of the hypersurface. Trajectories like that (by the way) are referred to in the literature as *ergodic*, and ergodicity (if you think about it) is clearly also the sort of thing one has in the back of one's mind when one speaks of trajectories wandering aimlessly and randomly and all over the place and in no particular direction.

Not *all* the trajectories can be like that, of course. Think (for example) of a gas in a perfectly rectangular container. And suppose that at a certain instant the N particles that make up that gas are all in a line (as shown in Figure 3.14), and suppose that their velocities at that moment are all equal, and parallel to one of the walls of the container. The trajectory of the system-point of a gas like that will be confined, forever, to a single tiny corner of its available energy hypersurface. The particles are going to be bouncing back and forth, just as they currently are, for all time.

But this is patently (in some sense, of which more later) a *very unusual case*. Alter the direction of the motion of even a *single* one of these particles, by even the *tiniest* angle, and wait around a while, and you will have a whole new ballgame. The *majority* of trajectories, the *vast* majority of trajectories, are (it would seem) going to *wander*. The vast majority of trajectories (it would seem) are going to be *ergodic*.

Let's make that a bit more precise. Call a point in phase space an *ergodic* point if the trajectory that point is sitting on—if the trajectory that point *determines*—is an ergodic trajectory. What the above considerations suggest, then, is that the area taken up by the *non*-ergodic points on any finite energy hypersurface of any sufficiently complicated Newtonian system is over-

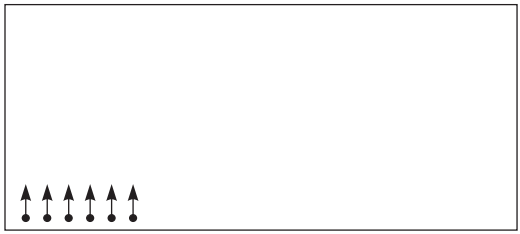


Figure 3.14

whelmingly tiny.[19] What the above considerations suggest (as a matter of fact) is that the area taken up by the non-ergodic points on any finite energy hypersurface of any sufficiently complicated Newtonian system (although there are invariably an infinite *number* of them) is *zero*.[20]

Good. Let's get back to our story. Consider a gas whose system-point is known to be located at a certain initial time in a certain particular non-equilibrium region of the sort of hypersurface depicted in Figure 3.13. And suppose that the trajectories on this hypersurface are typically *ergodic*. Then—since the overwhelming majority of the area of this hypersurface is taken up by its *equilibrium*-region—the typical trajectory on this hypersurface will spend the overwhelming majority of its *time* in that region. And so a system whose phase point is initially *outside* of the equilibrium region will typically *make its way there* before too long. And a system whose phase point is initially *within* the equilibrium region will typically *stay* there.

And so we have irreversibility yet again.

▲▲▲ None of this *proves* anything, of course—we have merely (remember) been making crude stabs; we have been making assumptions all over the place.

19. The *finiteness* of the energy hypersurface, by the way, is crucial here. To begin with, the very *definition* of ergodicity—as applied to trajectories—will patently *fail to make sense* if the area of the energy hypersurface on which the trajectory in question is located is infinite. But it's more than just that. It's that trajectories on infinite energy hypersurfaces are typically *not* going to have endless random-looking *twists and turns* in them. Think, for example, of the trajectories of a system consisting of N free particles, unconfined by any boxes, alone in infinite space.

20. Rigorous proofs of this sort of thing are hard to come by, but there are a few. It is now known, for example, that a system consisting of three hard spheres—three *billiard balls*, more or less—confined within a rectangular container, has this property.

But there can be no denying that these stabs have an enormously powerful cumulative suggestiveness. And the terminological resources for saying precisely *what it is* that they suggest are now (at long last) fully in place.

It goes like this.

Consider a true thermodynamical law, *any* true thermodynamical law, symmetric under time-reversal or not, to the effect that macrocondition *A* evolves under such-and-such external circumstances over such-and-such an interval into macrocondition *B*. What these stabs suggest (and suggesting this—mind you—is all these stabs are *for*, and so everything that's been said up to this point in this chapter is a ladder that can now be kicked away; everything that's been said up to this point in this chapter has now served its purpose by bringing us precisely *here*) is as follows: wherever such a law holds, it will be the case (that is, it will be a *logical consequence* of the *Newtonian particulate equations of motion*) that the overwhelming majority of the volume of the region of phase space associated with macrocondition *A* is taken up by microconditions which are sitting on trajectories which pass, deterministically, under the allotted circumstances, at the end of the allotted interval, through the region of phase space associated with macrocondition *B*.

And so if we are initially given only the information that macrocondition *A* obtains, and that the external circumstances are such-and-such, and if we suppose (in the absence of more detailed information) that the initial microcondition of the system in question is as likely to have been located in one part of the region of the phase space associated with macrocondition *A* as in another, if (that is) we suppose (in the absence of more detailed information) that the probability that the initial microcondition of the system in question was located in any particular *sub*region of the region of the phase space associated with macrocondition *A* is proportional to the *volume* of that subregion, and to nothing else, then it will follow that the probability of *A*'s evolving into *B*, under the allotted circumstances, over the allotted time, in the absence of more detailed information, is (just as the laws of thermodynamics dictate) overwhelmingly high.[21]

21. Note (by the way) that there will patently be any number of *other* ways of cutting any phase space up into chunks, into macroconditions, in terms of which there turn out *not to be* any simple and robust and more or less deterministic "laws of thermodynamics." The way our *senses* happen to have cut it up, then, is no accident.

And so there would seem (on the face of it) to be good reasons for suspecting that the laws of thermodynamics can in principle be deduced in their entirety, in all their irreversibility, from the Newtonian particulate equations of motion together with a single stunningly simple and eminently innocent-looking and manifestly time-reversal-symmetric postulate about statistics.

▲▲▲ It turns out (of course) that a hell of a lot about this, as it stands, is a hell of a lot too easy. And that's what much of the remainder of this chapter and much of the next one are going to be about.

2. THE NATURE OF THE POSTULATE ABOUT STATISTICS

I've been talking about the postulate about statistics up to now as if it more or less amounted to a stipulation that what you ought to suppose, for purposes of predicting a system's future behavior, if you are given only the information that the system initially satisfies X,[22] is that the system is as likely to be in any one of the microconditions compatible with X at the initial time in question as it is to be in any *other* one of the microconditions compatible with X at the initial time in question. That's more or less what the postulate amounts to (I think) in the imaginations of most physicists. And that (to be sure) has a supremely innocent ring to it. It sounds very much—when you first hear it—as if it is instructing you to do nothing more than attend very carefully to *what you mean*, to *what you are saying*, when you say that all you know of the system at the time in question is X. It sounds very much as if it is doing nothing more than reminding you that what you are saying when you say something like that is that X is the case at the time in question, and (moreover) that you have no more reason for believing that the system is in any particular one of the microconditions compatible with X at the time in question than you have for believing that it is in any *other* particular one of the microconditions compatible with X at the time in question, that (insofar as you know, at the time in question) nothing *favors* any particular one of those microconditions over any particular other one of them, that (in other words) the *probability* of any particular one of those microconditions obtaining at the time in ques-

22. If (for example) you are given only the information that the system is initially in a certain particular *macrocondition*.

tion, given the information you have, is *equal* to the probability of any particular other one of them obtaining at the time in question.

This is all wrong, however. And there is a technical reason that it's wrong, and there is a more fundamental (and less often rehearsed) one too.

The technical reason has to do with the fact that the sort of information we can actually *have* about physical systems—the sort that we can *get* (that is) by *measuring*—is invariably compatible with a *continuous infinity* of the system's *microconditions*.[23] And so the only way of assigning equal probability to all of those conditions at the time in question will be by assigning each and every one of them the probability *zero*. And *that* will of course tell us *nothing whatsoever* about how to make our *predictions*.

And so people took to doing something *else*—something that looked to them to be very much in the same *spirit*—*instead*. They abandoned the idea of assigning probabilities to individual microconditions, and took instead (of course) to stipulating that the probability assigned to any *finite region of the phase space* which is entirely compatible with X—under the epistemic circumstances described above—ought to be proportional to the continuous *measure* of the points *within* that region.

But there's a problem with that—or at any rate there's a problem with the thought that it's *innocent*—too. The problem is that there is in general an *infinity* of equally mathematically legitimate ways of *putting* measures on infinite sets of points. Think, for example, of the points on the real number line between 0 and 1. There is a way of putting measures on that set of points according to which the measure of the set of points between any two numbers a and b (with $a < 1$ and $b < 1$ and $b > a$) is $b - a$, and there is *another* way of putting measures on that set of points according to which the measure of the set of points between any two numbers a and b (with $a < 1$ and $b < 1$ and $b > a$) is $b^2 - a^2$. According to the first of those two formulae, there are "as many" points between 1 and ½ as there are between ½ and 0, and according to the *second* of those two formulae, there are *three times* "as many" points between 1 and ½ as there are between ½ and 0, and there turns out to be no way whatsoever (or at any rate none that anybody has yet dreamed up)

23. This follows from the fact that the totality of the possible microconditions of any Newtonian system invariably has the cardinality of the continuum, and that the accuracies of the measurements that we are able to perform are invariably *finite*.

of arguing that either one of these two formulae represents a truer or more natural or more compelling measure of the "number" or the "amount" or the "quantity" of points between *a* and *b* than the other one does.[24] And there are (moreover) an infinite number of *other* such possible measures on this interval as well, and this sort of thing (as I mentioned above) is a very general phenomenon.

And anyway, there's a much more fundamental problem. There's something completely *insane* (if you think about it) about the sort of explanation we have been imagining here. Forget about all the stuff in the last three paragraphs. Suppose there was no problem with the measures. Suppose that there were some unique and natural and well-defined way of expressing, by means of a distribution-function, the fact that "nothing in our epistemic situation favors any particular one of the microconditions compatible with X over any other particular one of them." So *what?* Can anybody seriously think that that would somehow *explain* the fact that the *actual microscopic conditions* of *actual thermodynamic systems* are *statistically distributed in the way that they are?* Can anybody seriously think that it is somehow *necessary*, that it is somehow *a priori*, that the particles that make up the material world must arrange themselves in accord with *what we know*, with *what we happen to have looked into?* Can anybody seriously think that our merely being *ignorant* of the exact microconditions of thermodynamic systems plays some part in *bringing it about*, in *making it the case*, that (say) *milk dissolves in coffee?* How could that *be?* What can all those guys have been *up* to? If probabilities have anything whatsoever to do with how things actually fall out in the *world* (after all), then knowing nothing whatsoever about a certain system other than X can in and of itself entail nothing whatsoever about the relative probabilities of that system's being in one or another of

24. This is (as a matter of fact) a long and not uninteresting story. There has been an entire tradition of attempts over the past hundred years or so to argue that (notwithstanding the undisputed formal mathematical infinity of measures) there will be features of every possible *physical situation*, or of every possible *epistemic relation* to a physical situation, which will dictate that there is exactly *one* way of putting measures on sets of *microconditions* that is (somehow) "natural" to that situation; and (moreover) that the appropriate sort of examination of the *symmetries* of that situation will invariably reveal *what way that is.* And this tradition is now (after much consideration) more or less universally acknowledged to be a failure. All of this is very nicely laid out, I think, in chapter 12 of a wonderful and seminal book of Bas van Fraassen's called *Laws and Symmetry* (New York: Oxford University Press, 1989).

the microconditions *compatible* with X; and if probabilities have *nothing* whatsoever to do with how things actually fall out in the world, then they can patently play no role whatsoever in explaining the behaviors of *actual physical systems*; and that would seem to be all the options there are to *choose* from!

Let's see where that leaves us. Certainly it does *not* follow merely from the fact that all we know of a certain system is X that the chance of that system's microcondition being located in any particular *sub*region of the region of the phase space that's *compatible* with X is proportional to the *volume* of that subregion. And yet it *does* seem to be some sort of a *fact*—or at any rate it seems to yield correct *predictions* to *suppose* that it is some sort of a fact—that the *percentage* of any *large collection* of randomly selected X-systems whose microconditions lie within any particular subregion of the X-region of the phase space will be more or less proportional to the familiarly defined *volume* of that subregion. And so the sort of fact that is must be an *empirical* one, a *contingent* one, a *scientific* one.

▲▲▲ But that can't be all there is to it, either. It turns out not to be quite *right*, it turns out to be not quite *true*, as a general matter, as an empirical matter, if you think it over, that the percentage of any large collection of randomly selected X-systems whose microconditions lie within any particular subregion of the X-region of the phase space is proportional to the volume of that subregion.

Suppose, for example, that X is the property of being an apartment that contains a spatula, and consider the large collection of such apartments on earth. If the percentage of any large collection of randomly selected X-systems whose microconditions lie within any particular subregion of the X-region of the phase space is as a general matter proportional to the volume of that subregion, then the ratio of the percentage of those apartments in which the spatula is in the kitchen drawer to the percentage of those apartments in which the spatula is (say) in the *bathtub* ought to be equal to the ratio of the amounts of *space* those two containers *take up*. But that's just not *right*. Spatulas (as a matter of fact) are hardly *ever* in bathtubs; or at any rate, they are *much* less often in bathtubs than they are in kitchen drawers; or at any rate, that's how it is on earth.

Or suppose that X is the property of being a glassful of water,[25] a glassful (more particularly) in which the average kinetic energy of the water molecules is well above the temperature at which water *freezes*. And consider the large collection of such glassfuls on earth. If the percentage of any large collection of randomly selected X-systems whose microconditions lie within any particular subregion of the X-region of the phase space is as a general matter proportional to the volume of that subregion, then the percentage of those glassfuls in which any appreciable amount of the water is *frozen* ought to be well within a millionth of a millionth of a millionth of a percent of zero. But we see glasses of water, sitting in *warm* rooms, with chunks of *ice* in them, *all the time*.

Maybe the thing to do is to *narrow down* the *antecedent* a bit; maybe it ought to be rewritten so as to refer not to *any* property *whatsoever* (which is how we've been writing it so far, which is what we've been meaning by the X) but only to the property of being in one or another particular *macrocondition*. Let the postulate read (then) that if a certain system is at present in a certain macrocondition *M*, then the probability that that system's *microcondition* currently lies within any particular *sub*region of the *M*-region of its phase space is proportional to the familiarly calculated *volume* of that subregion.[26]

Or something like that. That will do (at any rate) for the time being—but there are two points it will be well (for future reference) to keep in the back of one's mind.

The first is that this way of fixing things up is in a certain sense exceedingly *crude*. The trouble with the *original* postulate (remember) was that it seemed to be making false claims about (say) the locations of spatulas in apartments. And what we've done by way of *solving* that problem is simply to *rewrite* the postulate in such a way as to preclude it from making *any* claims about things like the locations of spatulas in apartments *at all*. And that would seem—or it *might* seem—to go a bit *too far*. There *do* appear to be such things in the world, after all, as robust statistical regularities about the

25. Tim Maudlin suggested this example to me.

26. And note that all this amounts (as well) to a contribution to our understanding of *what it is* to be a macrocondition *in the first place*—note (that is) that this is something to be kept in mind alongside of what we learned (say) in footnote 1 of Chapter 2, and in footnote 5 of this chapter.

locations of spatulas in apartments. And whatever such regularities there *are* will be rendered altogether *uncapturable* by our fundamental statistical postulate if we fix that postulate up as I am here proposing.

And note that (insofar as this *fixed-up* postulate *itself* is concerned) there must certainly be an infinity of *other* postulates that can serve all practical purposes just as well. The reason is that the familiarly calculated volume of the subregion of any *M*-region of the phase space of any thermodynamic system which is taken up by "abnormal" microconditions, microconditions (that is) that lead to *violations* of the laws of thermodynamics, is not merely *small* (which is what I have been at pains to emphasize about it so far) but also *scattered*, in unimaginably tiny clusters, more or less at random, all over the place, as illustrated (badly) in Figure 3.15. And so the percentage of the familiarly calculated volume of any regularly shaped and *not* unimaginably small subregion of the region of the phase space corresponding to whatever macrocondition the system in question happens to be in which is taken up by abnormal microconditions will be (to an extremely good approximation) the same as the percentage of the familiarly calculated volume of the region of the phase space corresponding to that macrocondition *as a whole* that abnormal microconditions take up. And so a uniform distribution, or even a reasonably *smooth* distribution,[27] over any regularly shaped and not unimaginably small subregion of the region of the phase space corresponding to whatever macrocondition the system in question happens to be in, will yield (to an extremely good approximation) the same probabilities of the system's being in this or that *other* macrocondition at this or that *future* time as does the above uniform distribution over the macrocondition as a whole.

3. WHAT ALL THIS IS ABOUT

Here (just by way of finishing up) are three quick sketches of a certain bad and influential train of thought about what the object of the science of statistical mechanics is.

a. *Entropies* (you will remember) are associated with *numbers of distinct arrangements*, or with *volumes of phase space*; and since any particular

27. That is, a distribution which varies only negligibly over distances of the order of the diameters of the unimaginably small subregions mentioned in the previous sentence.

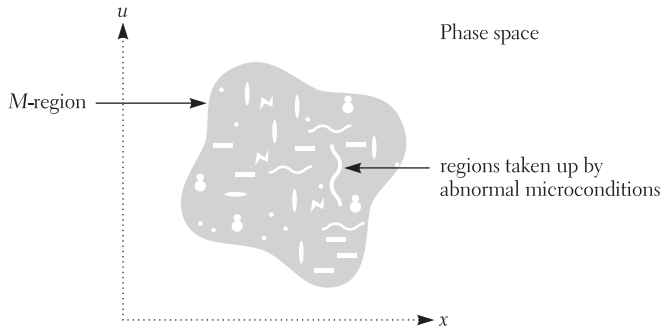


Figure 3.15

microcondition is necessarily compatible with exactly *one* arrangement,[28] since (to put it slightly differently) any particular microcondition occupies a volume of exactly *zero* in phase space, no two *microconditions* can *possibly* be associated with *different* entropies! And so the entropies of which we speak in thermodynamics, the entropies whose values can *change* with time, must somehow be associated not with individual systems but with *ensembles* of systems, or rather not with individual *conditions* but with volume-filling *probability-distributions* over conditions. What the entropies of which we speak in thermodynamics must *characterize* (to put it another way—a way which seems to link up with the talk outlined above about entropy and *information*) are not physical systems *per se* but rather *our knowledge* of such systems!

b. The briefest reflection on the fact that gasses are *collections of particles* (after the manner of the Maxwell's demon story, say) reveals that it cannot possibly be the case that individual systems will, as an invariable rule, monotonically evolve in any particular direction—that it cannot possibly be the case that individual systems will as an invariable rule evolve monotonically toward any particular *condition of equilibrium*. By contrast, it seems not at *all* implausible that *infinite ensembles* of systems (or rather, *probability-distributions* over *infinite collections of microconditions*) might be shown to evolve that way. And so it must be such *ensembles*, such *probability-distribu-*

28. And note that this will be the case no matter *how* we decide to carve up the phase space into macroconditions.

tions, and not individual systems, to which the laws and the predicates of thermodynamics actually, properly, apply.

c. To speak of something being in "equilibrium" is surely (among other things) to assert that its properties *do not vary with time*. But the same sorts of very elementary reflections that I alluded to in the last paragraph will reveal that there can certainly not be any individual microconditions of a gas that have this property. It turns out to be easy to show, by contrast, that there *is* a *probability-distribution* over the possible microconditions of any given system, subject to any given set of fixed gross constraints, that *does*.[29] And so "equilibrium," too, must be a predicate not of individual gasses but of *ensembles* of them; not of microconditions but of distributions *over* microconditions.

And now it becomes a matter of some urgency (since there is, after all, only *one* macroscopic physical condition of equilibrium corresponding to any given set of gross constraints) to show that there are no *other* probability-distributions that have this property. And it turns out to be possible to prove that if the laws of the motion of some particular Newtonian-mechanical system are *ergodic*, then there *aren't*[30] (the idea is roughly that the ergodicity of the laws of motion will entail that distributions *other* than the one that's uniform, on the standard measure, over the entire accessible region of the phase space—distributions which, say, are confined to some

29. The distribution in question here turns out to be the one that's *uniform* (relative to the standard measure) over the *complete set* of the microconditions which are *compatible* with those constraints. The demonstration (as I said) is easy. Note (to begin with) that for any point *P* in the phase space of any Newtonian-mechanical system, there will invariably be some unique *other* point in that space (let's call it the "*N*-second-evolution" of *P*) at which a system initially located at *P* will *end up*—in accord with the deterministic equations of motion—*N* seconds *down the line*. Consider the *N*-second-evolutions of every one of the points in some arbitrary *region* of the phase space of some such system. Those *N-second-evolutions* will (of course) form a region of the phase space too. And there is a very general and very beautiful theorem of Newtonian mechanics (which we will prove, as a matter of fact, in the next chapter) to the effect that although those two regions can differ arbitrarily in shape and location, they can*not* differ *at all*, for *any* value of *N*, in *volume*. And so the *N*-second-evolution of the region occupied by the complete set of the microconditions of any such system which is compatible with any particular fixed gross constraints can evolve into nothing (if you think about it) other than precisely *itself*.

30. Or rather, it turns out to be possible to prove that if the laws of the motion of some particular Newtonian-mechanical system are ergodic, then the only other probability-distributions with this property, if there are any, must assign non-zero probabilities to sets of microconditions whose standard *measures*, whose familiarly calculated *volumes*, *are* zero, and there turn out to be all sorts of reasons (or so they say) that a distribution like *that* cannot possibly represent *any macrocondition whatsoever*.

smaller *subregion*—will invariably end up leaking all over the place). And so the business of concocting rigorous proofs of the ergodicity of this or that set of Newtonian laws of motion is all of a sudden of the utmost foundational importance.

▲▲▲ *All* of this, however, is sheer madness. Let's try to keep our heads on. The sort of entropy we are attempting to get to the *bottom* of here, remember, is the entropy we ran into in *thermodynamics*. And thermodynamical entropy is patently an attribute of *individual systems*. And attributes of individual systems can patently be nothing other than attributes of their *individual microconditions*. And the way to *calculate* the entropy of an individual microcondition (if *that's* what the trouble is supposed to be) is patently to calculate the number of *arrangements* compatible with the *macrocondition* to which that microcondition *belongs*.[31] And as to the time-invariance of the condition of equilibrium and the monotonicity of the approach *to* equilibrium, the thing to say about *them* is just that they turn out not to be quite *true*. And the prodigious effort that has over the years been poured into rigorous proofs of *ergodicity* is nothing more nor less—from the standpoint of the foundations of statistical mechanics—than a waste of time.[32]

31. This sort of an attitude more or less goes back (I think) to Boltzmann, and it has been defended in recent years, and with considerable eloquence, by (among others) Shelly Goldstein and Joel Lebowitz.

32. Mind you, the effort has certainly produced beautiful mathematics. And insofar as the project of a statistical-mechanical explanation of the laws of thermodynamics is concerned, it is certainly *suggestive*, it is certainly *welcome*, that the laws of the motions of the phase points of a wide variety of thermodynamic systems appear, or appear *in some approximation*, to be ergodic. But neither the ergodicity nor the *approximate* ergodicity of those sorts of systems is necessary for the success of that project, and neither of them would suffice; and the *difference* between them, insofar as this particular project is concerned, is of no importance whatsoever.

THE REVERSIBILITY OBJECTIONS AND THE PAST-HYPOTHESIS

Let's see where we are.

It would appear that the tension we started out with has somehow magically *evaporated*. It would appear—if everything goes as expected, if everything goes as the results of the previous chapter suggest—that the laws of thermodynamics can in principle be derived, in their entirety, in all their irreversibility, by pure logical deduction, from nothing over and above the Newtonian laws of motion and the postulate about statistics.

▲▲▲ But this *absolutely cannot be*. The laws of thermodynamics (once again) have a temporal *direction* in them; and there is patently *no* such direction, and there is patently nothing capable of *picking out* such a direction, anywhere in the Newtonian laws of motion; and there is patently no such direction, and there is patently nothing capable of picking out such a direction, anywhere in the postulate about statistics.[1]

This is worth rubbing in some. And there are a pair of century-old and perennially popular ways of doing that, and they are referred to in the literature as the *reversibility objections*.

The first objection—the objection of Zermello and Loschmidt, the objection I laid out (more or less) in Chapter 1—is that it would seem to follow from the invariance under time-reversal of the laws of motion and the postu-

1. The Newtonian laws of motion (as was pointed out at considerable length in Chapter 1) make no distinction whatsoever between past and future; and insofar as the postulate about *statistics* is concerned, the question of such a distinction cannot even arise. The postulate about statistics (remember) isn't about relationships *between* times *at all*; what it says (rather) is that a certain relationship obtains, at any *single* time, between the macrostate of the system in question and the probability of its being in some particular *microstate*.

late about statistics that entropy-*decreasing* processes can be no less natural or familiar or statistically common in the world than entropy-*increasing* ones.

Let's make that completely explicit. Suppose that a certain macrocondition *A* evolves, irreversibly, over *n* seconds, as a matter of thermodynamic law, into some other macrocondition *B*; and suppose that macrocondition *B* evolves, over *n* seconds, as a matter of thermodynamic law, into some *third* macrocondition *C*—*A* and *B* and *C* might (say) be macroconditions of an isolated *warm room*, in the first of which there is a block of ice, and in the second of which there is a half-melted block of ice and a puddle, and in the third of which there is no ice at all and a bigger puddle. And call a microcondition *m′* the *n-second-evolution* of another microcondition *m* (as I began to do in footnote 29 of the previous chapter) if the Newtonian laws of motion entail that *m* evolves, deterministically, over *n* seconds, into *m′*. And now consider any one of the microconditions of the system in question which is compatible with *C* and which also is an *n-second-evolution* of a microcondition compatible with *B* and which also is a 2*n*-second-evolution of a microcondition compatible with *A*. Take the *velocity-reverse* of that condition (which is to say, take the microcondition you get by starting with the one we just picked out—the one compatible with *C*—and reversing the velocities of all the particles, and leaving everything else unchanged) and let it evolve for *n* seconds into the future. What that will give you, by the time-reversal symmetry of the Newtonian laws of motion, is the velocity-reverse of a microcondition compatible with *B*. Let it evolve *n* seconds *more* into the future. What *that* will give you, by the time-reversal symmetry of the Newtonian laws of motion, is the velocity-reverse of a microcondition compatible with *A*. And note that the velocity-reverse of any microcondition compatible with *A* will *itself* be compatible with *A*, and that the velocity-reverse of any microcondition with *B* will itself be compatible with *B*, and that the velocity-reverse of any microcondition compatible with *C* will itself be compatible with *C* (nothing about what it is to be a warm room with a puddle or a block of ice or a combination of the two in it, after all, entails anything whatsoever about the *directions* in which any of the *individual molecules* in that room happen to be *moving*). And so for every single trajectory which is in accord with the laws and which carries you from *A* to *B* to *C*, there will necessarily be exactly one which is in accord with the laws and which carries you,

in the same amount of time, from *C* to *B* to *A*. And note (moreover) that the familiarly calculated *volume* of the region of phase space taken up by the velocity-reverses of any infinite set of microconditions will necessarily be *equal* to the familiarly calculated volume of the region taken up by that set *itself*. And so, given only the information that a certain room is at present (say) in macrocondition *B*, the postulate about statistics will entail (if you think about it) that the room is *exactly* as likely to be on its way from *C* to *A* as it is to be on its way from *A* to *C*.

The second objection is still more powerful—and more illuminating. It will take some setting up, however.

Think (to begin with) of a certain continuous set of points, a certain region, a certain *blob*, in phase space. And consider the set of *n-second evolutions* of *every one of the points* in that blob. That latter set will patently constitute *another* blob,[2] which will presumably have a different shape, and which will presumably be located in a different part of the phase space, from the first. But there is a theorem of Liouville to the effect that the familiarly calculated *volumes* of those two blobs (whatever the value of *n* is) will necessarily be *identical*.[3] The flows of points in phase space (that is) have

2. It will be a *single* blob rather than a *number* of them, by the way, because it is a property of the Newtonian laws of motion that the microproperties of any isolated system *n seconds* from now are invariably a *continuous* function of its microproperties *now*, at all times, and for any value of *n*.

3. The proof goes like this:

Every system describable by classical mechanics (which includes—among many other sorts of things—collections of point particles, interacting by means of separation-dependent interparticle forces, such as we have been concerned with here) can be uniquely associated with one or another so-called *Hamiltonian* function *H*. For systems consisting of particles, *H* is invariably a function of the particles' individual *positions*, and of their individual *momenta*, and of *nothing else*; and *H* invariably takes the form of the total *kinetic* energy of the system plus its total *potential* energy; and it happens (and this, of course, is why these functions are of interest in the first place) that the equations of the motion of any such system can invariably be written down, in an especially simple and elegant way, in *terms* of its Hamiltonian. More particularly, it happens that for every such system

$$dx_i/dt = dH/dp_i \tag{4.1}$$

and

$$dp_i/dt = -dH/dx_i. \tag{4.2}$$

Consider, for example, a single free particle, which is free to move about only in a single

precisely the mathematical structure of the *currents* in an *incompressible fluid*.[4]

And there is a technique, due to Poincaré, for parlaying this theorem into *another* one, a substantially more *striking* one, according to which any classical system which is confined to any finite region of its phase space (any classical system which, say, is *in a box*, and which has a certain definite, finite,

---

spatial dimension. For a particle like that (since its potential energy is always, by definition, zero),

$$H = p^2/2m, \tag{4.3}$$

and so it will follow from equations (4.1) and (4.2), respectively, that

$$dx/dt = p/m \quad \text{and} \quad dp/dt = 0, \tag{4.4}$$

which (of course) precisely describes the *motion* of a particle like that.

For cases of *multi*particle systems, with interparticle *forces* acting between them, which are free to move around in a *three*-dimensional physical space, the Hamiltonian will take the general form

$$H = p_i/2m_i + V(x_1 \ldots x_{3N}), \tag{4.5}$$

where N is the number of particles. And *that* will entail (via (4.1) and (4.2)) that

$$dx_i/dt = p_i/2m_i \quad \text{and} \quad dp_i/dt = dV/dx_i, \tag{4.6}$$

which, taken together, amounts (in a slightly different notation) to precisely $F = ma$.

Good. Now, let $f_k(x_1 \ldots x_{3N}, p_1 \ldots p_{3N})$ (where k varies from 1 to 6N) be what you might call the *flow*-field (or f-field) in phase space, by which I mean:

$$\begin{aligned} &f_1(x_1 \ldots x_{3N}, p_1 \ldots p_{3N}) = dx_1/dt_{[x_1 \ldots p_{3N}]} \\ &\vdots \\ &f_{6N}(x_1 \ldots x_{3N}, p_1 \ldots p_{3N}) = dp_{3N}/dt_{[x_1 \ldots p_{3N}]} \end{aligned} \tag{4.7}$$

It will follow from (4.1) and (4.2) (if you just write it all down) that the 6N-dimensional *divergence* of $f_k(x_1 \ldots x_{3N}, p_1 \ldots p_{3N})$, which is defined as

$$df_1(x_1 \ldots x_{3N}, p_1 \ldots p_{3N})/dx_1 + df_2(x_1 \ldots x_{3N}, p_1 \ldots p_{3N})/dx_2$$

$$+ \ldots + df_{6N}(x_1 \ldots x_{3N}, p_1 \ldots p_{3N})/dp_{3N},$$

which is (as it were) the *net outflow* of the f-field from any particular point in the phase space, is always, everywhere, *zero*.

And there is a famous theorem of C. F. Gauss to the effect that the net flow of any vector field outward across the boundary of any *finite region* is equal to the integral of the *divergence* of that field throughout the *interior* of that region. And note that the net flow of $f_k(x_1 \ldots x_{3N}, p_1 \ldots p_{3N})$ outward across any such boundary will be equal to the rate at which the *volume* of the set of *points* in that boundary is *increasing*.

QED.

4. And this makes good sense, if you think about it; it has something intuitively to do with the *determinism* of the classical equations of motion, with the fact that no two trajectories in the phase space ever *merge*, and that no single one ever *splits*.

total energy) will in the long run *invariably return to its initial conditions*, or (at any rate) *arbitrarily close* to its initial conditions.

The proof goes like this:

Consider some particular blob in phase space called g_0, and suppose that the volume of g_0 is w_0. The blob which is the *n-second evolution* of g_0 is called g_n, and *its* volume is called w_n, and it follows from Liouville's theorem that $w_0 = w_n$.

Now, think of the much *bigger* blob which is made of the union of *all* the points contained in *any* of the blobs g_t for $0 < t < \infty$. Call that big blob G_0, and call its volume W_0; and let G_T for any $0 < T < \infty$ represent the big blob composed of the union of all the points contained in any of the little blobs g_t for $T < t < \infty$; and let W_T represent the volume of G_T; and note that the volumes of all the G_T will clearly be *finite* (since, by hypothesis, the volume of the *entire available phase space* is finite here); and note that G_0 will clearly *contain* all the G_T.

And now switch gears. G_0, whatever *else* it is, whatever methods were used to *construct* it, is patently itself a *blob* in *phase space*. And we can perfectly well inquire, if we wish, what (say) the *T-second evolution* of G_0 is. And the *T*-second evolution of G_0 is patently G_T. And so (by Liouville's theorem) $W_0 = W_T$. And so (since G_0 *contains* G_T) it must be that G_0 and G_T are as a matter of fact (modulo, at most, a set of points whose cumulative measure is *zero*) *the same blob!* And so G_T, for any value of *T*, must contain (among other things) g_0. And remember that G_T is entirely made up of *future-evolutions* of g_0. And so it follows that all the points in g_0 are among the *future-evolutions* of the points in g_0!

And what that means is (of course) that all the points in g_0 (except, perhaps, for a set of measure zero of them) are sitting on trajectories which passed through g_0 at some *earlier* time. And what *that means* (in light of the invariance under time-reversal of the Newtonian equations of motion) is that all the points in g_0 (except, perhaps, for a set of measure zero), will pass through there *again*. And note that all this will be the case no matter how small g_0 may happen to be.

And so any classical-mechanical system which is confined to a finite region of phase space will sooner or later loop back around, with probability one, arbitrarily close to *all* the points in that space which it ever occupied, and then again, and then again, and so on, ad infinitum.

And this is an astounding result. Consider what it implies:

1. A drop of ink, which is dropped into a bowl of clear water, in a closed room, will eventually, with probability one, re-collect within the water and hop back out.
2. An egg, dropped on the floor from a considerable height in a closed room, will eventually, with probability one, reassemble itself perfectly on the floor and then hop up off it.
3. Two gasses of different temperatures in thermal contact with each other will eventually, with probability one, return to their original, different, temperatures.

.
.
.

N. All irreversible processes that occur within closed, finite volumes of phase space will eventually, with probability one, and then again and again and again, *reverse* themselves.

And so it ineluctably follows from the Newtonian laws of motion and the best version we currently have of the postulate about statistics that (1) the entropy of any thermodynamic system in any macroscopic condition is exactly as likely to be *decreasing* as it is to be *increasing*; and that (2) the number of entropy-decreasing segments and the number of entropy-increasing segments of the full history of any individual thermodynamic system which is permanently isolated and permanently confined to some particular finite region of phase space will, with probability one, be *equal*. And so things are absolutely and positively *not* as I said they appeared in the third sentence of this chapter.

What the hell was the content of the insight of Boltzmann and Gibbs, then? Nothing?

Let's slow down a bit. Let's think it through more carefully.[5]

What they saw, or what (at least) they were able to make plausible, was that the Newtonian laws of motion and the postulate about statistics (or some-

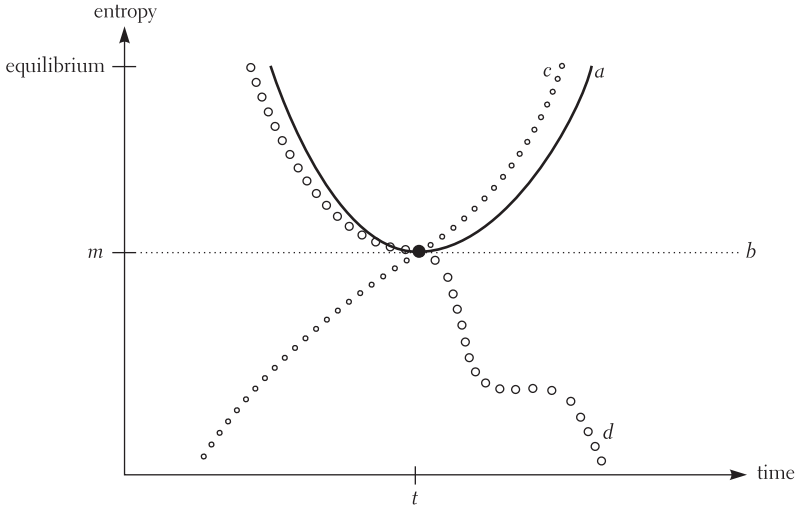
5. And these next few paragraphs will more or less rehearse, insofar as I am familiar with it, what Boltzmann and Gibbs *themselves* had to say, at the time, by way of responding to the two objections we have just been discussing.

thing *like* the postulate about statistics, but of that more in a minute) entail that the *overwhelming majority* of the trajectories passing through any particular non-maximal-entropy macrocondition *increase their entropies toward the future*; that (more precisely) the *measure*, or rather the *standard* measure, of the trajectories passing through any particular non-maximal-entropy macrocondition which increase their entropies toward the future is overwhelmingly *larger* than the standard measure of those that *don't*.

And this (if you think about it) is something a good bit *weaker* than the proposition that gets alluded to in the third sentence of this chapter. This—as it stands—is perfectly compatible with the time-reversal-symmetry of the Newtonian laws of motion and the postulate about statistics, and (more particularly) it is perfectly compatible with the claim that entropy-increasing trajectory segments are no more plentiful, either in general (which is to say, within the entire set of physically possible trajectories) or even in the history of any particular single, isolated, bounded, thermodynamic system, than entropy-*decreasing* trajectory segments are. To say that the overwhelming majority of the trajectories passing through any particular non-maximal-entropy macrocondition increase their entropies toward the future is (after all) certainly not to *deny* that the overwhelming majority of the trajectories passing through any particular non-maximal-entropy macrocondition also increase their entropies toward the *past!*

What Boltzmann and Gibbs saw, then, was apparently that the Newtonian laws of motion and the postulate about statistics (or something like it, of which more in a minute) entail that the overwhelming majority of the trajectories passing through any particular non-maximal-entropy macrocondition must just then be in the process of *turning around*; that (as illustrated in Figure 4.1) every non-maximal-entropy macrocondition of a thermodynamic system represents a local entropy *minimum* of the overwhelming majority of the physically possible trajectories that pass through it.

Let's get in a bit deeper here. Let's see how that can be. Consider a system which the laws of thermodynamics say will evolve from a certain relatively low-entropy macrocondition *A* through a certain higher-entropy macrocondition *B* to a certain equilibrium macrocondition *C*. If statistical mechanics is to be able to do its *job*, then, the overwhelming majority of the physically possible trajectories passing through *A* are going to have to pass, subsequently, through *B*. And yet Boltzmann and Gibbs, as we are currently



The overwhelming majority of trajectories which pass, at *t*, through a non-equilibrium macrocondition *M*, whose entropy is *m*, will look like *a*. A vastly tinier fraction will look like *b* or *c* or *d* or what have you.

Figure 4.1

reading them, are going to be committed to the proposition that *B* represents a local entropy *minimum* of the overwhelming majority of the physically possible trajectories that pass through it; which is to say that Boltzmann and Gibbs, as we are currently reading them, are going to be committed to the proposition that only an unimaginably tiny *minority* of the physically possible trajectories passing through *B* have recently passed through *A*. And all of this can only be *consistent* in the event that (as it were) the *total* number of physically possible trajectories passing through *B* is overwhelmingly larger than the total number of physically possible trajectories passing through *A*; in the event (that is) that the total number of distinct microconditions *compatible* with *B* is overwhelmingly larger than the total number of distinct microconditions compatible with *A*. But (of course) it *is*. That, after all, is *just what it amounts to*, in the language of statistical mechanics, to say that *B* has higher *entropy* than *A* does!

▲▲▲ Good.

But now there are patently new concerns. And quite serious ones.

The above considerations make it abundantly clear (to begin with) that even the most recent and most sophisticated of our attempts at cooking up a postulate about statistics can't possibly be quite right, because it turns out to be radically *incompatible* with what we have just now come to suspect must be true of the Newtonian laws of motion.

The trouble (more particularly) is that in the event that the postulate is true of a certain ensemble of thermodynamic systems at one time, it will in general *not* be true of that ensemble of systems at any *later* time.

Suppose (for example) that at $t = 0$ we place a large ensemble of fully unmelted ice cubes in an equally large ensemble of warm rooms. And suppose that this composite ensemble is initially well described by the most recent version of our postulate about statistics. Well, that postulate, together with the Newtonian laws of motion, is going to entail (if all goes as we now hope) that the overwhelming majority of these ice cubes will be partly melted at (say) $t = (5$ minutes). But suppose that we were now to apply the postulate *again,* this time to the macrocondition of the composite ensemble at $t = (5$ *minutes).* Well, what *that* will entail, together with the Newtonian laws of motion, as we have just been discussing, is that back at $t = 0$ the overwhelming majority of the ice cubes were *more melted still!*[6] And that (of course) flatly contradicts one of the postulates we *started out* with! And so it turns out to be a straightforward mathematical impossibility (if the Newtonian equations of motion have the sorts of consequences, for systems like these, that Boltzmann and Gibbs suggest they do)[7] for there to be any ensemble of melting ice cubes which is well described, throughout its evolution, by this most recent version of our postulate about statistics. And as a matter of fact (if you think it over for a minute) there turns out not even to be a *single* possible macrocondition of a classical system—not even its *equilibrium* condition—which has the property that if *it* is the initial macrocondition of the isolated system in question, and if the static probability rule initially holds,

6. That is, what the most recent version of our postulate about statistics will entail, when applied to the macroconditions of the composite ensemble of the systems we have been talking about at $t = (5$ minutes), if the Newtonian laws of motion have the property we have lately come to think they must, is that the proportion of those systems whose *microconditions* at $t = (5$ minutes) are 5-minute evolutions of microconditions in which the ice is completely unmelted is overwhelmingly *small.*

7. Which is that every non-maximal-entropy macrocondition represents a local entropy *minimum* of the overwhelming majority of the physically possible trajectories that pass through it.

then the dynamics will guarantee that the rule will continue to hold in the future.[8]

Moreover, all questions of logical compatibility aside, the considerations of the last few pages make it clear that our latest version of the postulate about statistics is explicitly, empirically, *false* (of large systems of half-melted

8. It will turn out to be worth our while, for future reference, to take a few sentences to go through this business of compatibility (or the lack of it) between a dynamical law and a static probability rule in full generality.

Consider (then) any static probability rule (which is to say, any rule concerning temporal *coincidences* of properties, rather than their temporal *sequences*) which gives the probability that X_j will be instantiated, at any particular time, given that Z_i is instantiated then (in the statistical mechanics we have been working through here, for example, the Z_i are clearly the macroconditions, and the X_j are the microconditions).

Now, a probability rule of this sort is said to be *incompatible* with a law of dynamical evolution in the event that the dynamics can take us from circumstances which are in accord with the probability rule to circumstances which are *not* in accord with it.

Let's spell this out in somewhat more detail. The static probability rule (whatever particular form it takes) will associate some particular X-distribution with each particular Z_i. Consider some initial Z_i – (X-distribution) pair which is in accord with some particular static rule. Evolve this initial situation into the future by means of the dynamical laws; this will yield a probability-distribution over the various possible Z-values at the future time in question, and it will also yield a probability-distribution over the various possible X-values at the future time in question, and it will *also* yield a collection of *conditional* probability-distributions over the various possible X-values *given* one or another of those Z-values whose overall probabilities, at the future time in question, are non-zero. *If, for any initial value of* Z, *any of the above-mentioned conditional probability-distributions is not in accord with the static rule in question, then that rule is said to be incompatible with the dynamics with which those conditional distributions were calculated.*

And so (once again) it will now follow from the argument given in the text that if what we have recently come to suspect of the dynamical laws of Newtonian mechanics is true, then the dynamical laws of Newtonian mechanics are *not* compatible with the stipulation that the appropriate probability-distribution for microconditions is (at all times) the one that's uniform over the present macrocondition.

And there is a famous deterministic hidden-variable version of non-relativistic quantum mechanics (which is due to David Bohm; and of which we shall be talking a good deal more, in another context, near the end of this book) which will make for an instructive comparison here. Never mind (for the moment) what a theory with a name like that might be *for*. What I want to focus on at present is just its formal *structure*.

The world (according to Bohm's theory) has two sorts of physical objects in it: material particles and something called *wave-functions*. And all the particles and all the wave-functions always evolve (as I said above) in accord with thoroughly *deterministic* dynamical laws. The way it works is this: the wave-functions have their own completely autonomous and thoroughly deterministic laws of evolution, and proceed on their various ways, completely *oblivious* to, completely *independent* of, the behaviors of the material particles. The *particles*, by contrast, are *very much* affected by the *wave-functions*. It is the wave-functions (as a matter of fact) that more or less single-handedly *carry the particles along*, as corks (say) are carried along on a river. And the laws of that carrying are also fully deterministic, so that if the wave-functions and the positions of the particles are all given at some initial time, their values at any *later* time can in principle be

ice cubes, say) *in our world*, since such systems almost invariably *do* (as a matter of fact) originate from systems of fully unmelted cubes.[9]

And finally (and this is really just a restatement, in somewhat grander terms, of the *previous* point), this latest version of the statistical postulate, if applied in the present, is *flatly inconsistent* with what we take to be true, with what we *remember*, with what is *recorded* (that, for example, everybody was younger, that ice was less melted, and so on), of the *past*. The postulate, as we have it now, if applied in the present, will entail (after all) that throughout the past the entropy of the universe has constantly been *decreasing*.

▲▲▲ Let's think about what we might be able to *do* about all that.

Let me remind you, to begin with, of something we learned in the previous chapter. There I had taken to referring to those regions of the phase space of a thermodynamic system which correspond to that system's possible *macrostates* (remember) as "*M*-regions." And I remarked at a certain point that the familiarly calculated volume of the subregion of any *M*-region of the phase space of any thermodynamic system which is taken up by microstates that lead to decreases in entropy toward the future is not merely *small* (which is what the bulk of that chapter was all about) but also *scattered*, in

---

determined from the dynamical law of the evolutions of the wave-functions, and the dynamical law whereby the wave-functions carry the *particles* along, with complete certainty.

And it happens that Bohm's theory also has a *non*-dynamical law in it, which takes the form of precisely the sort of static probability rule that we have just been thinking about, which consists of an algorithm for calculating the conditional probability that the material particles in the world are (at a certain time) located in such-and-such a set of positions, given that the *wave-functions* of the world are (at the time in question) in such-and-such a *configuration*. And so the configurations of the wave-functions are the Z_i's here, and the positions of the material particles are the X_j's. And it happens (and this is the twist) that the truth of the *Bohmian* static probability rule, at *all* times, *is fully compatible with the Bohmian dynamical laws*. Indeed (and this just comes to the same thing, given that the Bohmian dynamical laws are fully deterministic), the truth of that rule at any *one* time, together with the laws of dynamics, turns out to *entail* its truth at any other!

And so it is by no means a matter of general necessity that dynamical laws and static probability rules—even of the sort that you run into in rich and serious and fully developed scientific theories—must necessarily *collide* with each other. It is (rather) an unfortunate peculiarity of the formulations we have thus far been able to cook up of *classical statistical mechanics*.

9. Here (by the way), and everywhere else in statistical mechanics, and (as a matter of fact) everywhere *simpliciter*, it seems to me to be a great help, it seems to me to spare one an enormous amount of confusion, to be thinking of probabilities as *supervening* in one way or another on the *non*-probabilistic *facts of the world*, to be thinking of them (that is) as having something or other to do, by definition, with *actual frequencies*.

unimaginably tiny clusters, more or less at random, all over the place (and this was very crudely illustrated in Figure 3.15, which appears on page 68). And what we have learned so far in the *present* chapter will then suggest that *precisely the same thing* must be true of the subregion of any *M*-region of the phase space of any thermodynamic system which is taken up by microstates that lead to decreases in entropy toward the *past*. And (moreover) there is patently no reason at all that those two subregions of any particular *M*-region of the phase space of any particular thermodynamic system[10] should have any tendency whatsoever to be *aligned* or to be *correlated* or to be otherwise *matched up* with each other. And so the percentage of the familiarly calculated volume of either *one* of those tiny subregions which is taken up with its *intersection* with the *other* one of those tiny subregions will very likely be very nearly the same as the percentage of the familiarly calculated volume of the region of the phase space corresponding to that macrocondition *as a whole* that the latter subregion takes up.[11]

And so the apparent time-invariant rightness of the future thermodynamic predictions of the standard uniform-over-the-present-macrocondition distribution need not force us into the self-contradictory posit that that distribution always actually *obtains!* There might be ways of making it suffice if that distribution were to obtain only *once*, and *not now*.

Consider, for example, the following system (which is pictured in Figure 4.2): a long line of glasses is arrayed underneath a triangular pinballish device. The glasses all contain warm water, and at present a number of them contain half-melted cubes of ice.

The standard uniform-over-the-present-macrocondition distribution and the laws of motion (if what we have lately come to believe about them is true) will entail that the ice will with overwhelmingly high probability be fully melted five minutes from now, and that is of course very much in accord with what we have learned to expect from past experience, and with what the second law of thermodynamics requires. But that same distribution

10. That is, the subregion that's taken up by microstates that lead to decreases in entropy toward the *future* and the subregion that's taken up by microstates that lead to decreases in entropy toward the *past*.

11. Note that it's the "randomness" and the "lack of alignment" of the subregions we're thinking about here that are fulfilling the function of the "smoothness" and the "regularity" of the distributions we were thinking of in connection with Figure 15 of Chapter 3.

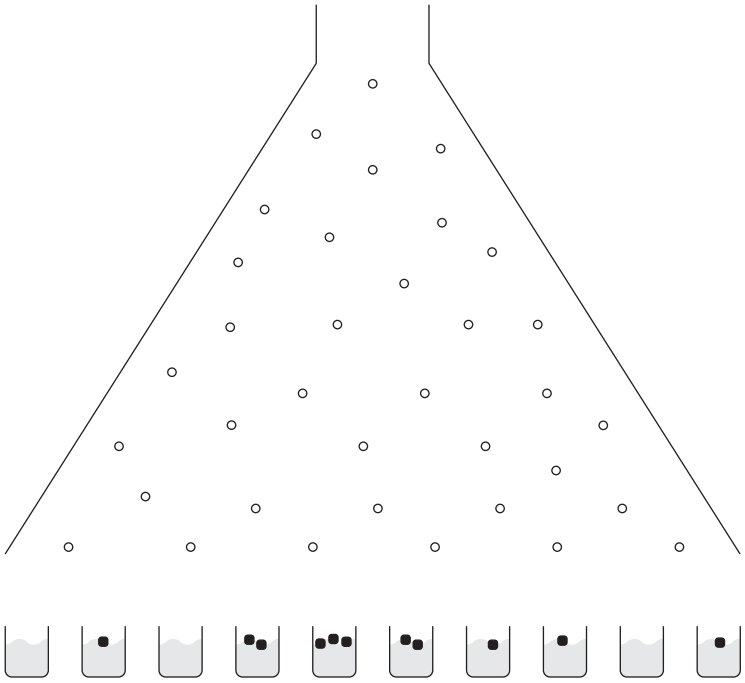


Figure 4.2

(together with the laws of motion) *also* entails, with *equally* overwhelming probability, that the cubes were fully melted five minutes *ago*, and (more generally) that the entropy of these systems has been *decreasing* over that time, and *this* (of course) runs counter to the laws of thermodynamics, and to our past experience with systems of this general type, and (let's suppose) to our *specific memory* of the past five minutes of the history of this *particular system* as well.

Here's how to remedy that: posit (in *accord* with our memory) that five minutes ago the glasses in question had fully unmelted ice cubes in them, and that the microcondition-probability was uniform—on the standard measure—over the macrocondition *then*. Relative to *this* posit, it will be overwhelmingly probable that the present state is what it is, and that the future will be what all our previous experience of such systems and the summary of that experience in the second law lead us to *expect* it to be, and that the past five minutes were what we *remember* them to be.

But this *new* posit is still plainly going to get everything about the period *prior* to five minutes ago *wrong:* it will dictate, for example, that *ten* minutes ago the glasses in question almost certainly contained half-melted ice, which subsequently (in violation of the second law) became fully *un*melted ice. Here's how to remedy *that:* posit (and suppose, once again, that this is in accord with our memory) that ten minutes ago the ice was all collected at the top of the pinballish device. Suppose (moreover) that the average temperature of this macrocondition is somewhat *lower* than the average temperature of the macrocondition we referred to in the *previous* posit (the temperature difference here will correspond to the energy gained as the ice falls down into the glasses). And posit that the microprobability-distribution is uniform—on the standard measure—over that ten-minutes-ago macrocondition.

The details of this case are going to be interestingly different from those of the one above. It is (to begin with) certainly *not* the case that this last posit will make either the present macrocondition or the five-minutes-ago macrocondition overwhelmingly probable: this posit (as a matter of fact) will make *no* particular present or five-minutes-ago macrocondition overwhelmingly probable. What it will do (rather) is to make certain prominent thermodynamic *features* of the present and five-minutes-ago macroconditions overwhelmingly probable (their average temperatures, for example, and the degree to which what ice there is in them is *melted*, and so on), but it will clearly assign similar probabilities to a rather wide variety of quite *distinct* five-minutes-ago macroconditions (macroconditions associated with the ice cubes having landed in quite different sets of glasses, for example). What we *have*, though, in this last posit, and what we were *lacking* in the previous one, is a probability-distribution relative to which what we remember of the entirety of the last ten minutes, and what we know of the present, and what we expect of the future, is (you might say) *typical*.[12] And so what we have in the conjunction of this last posit with the laws of classical mechanics is (if all goes as expected) a fully satisfactory probabilistic *theory* of the history of this system beginning ten minutes ago.

12. What we have in this last posit (that is) is a probability-distribution relative to which a certain highly restricted set of sequences of macrostates—a set which happens to include what we remember of the entirety of the last ten minutes, and what we know of the present, and what we expect of the future—is overwhelmingly more probable than any *other* such sequence.

But of course this last posit is *itself* going to go bad *prior* to ten minutes ago, and by now it will be perfectly clear that *all* such posits are bound to fail—unless they concern nothing less than the *entirety* of the universe at nothing later than its *beginning*.[13]

That's what the statistical posit is going to have to be about, then. And if the project of statistical mechanics is on anything remotely like the right track, then, when all the data are in, the initial macrocondition of the universe had better turn out to be one relative to which—on the standard uniform probability-distribution over microconditions—what we think we know of the history of the world, and what we expect of its future, is *typical*. And what seems to me to have been the achievement of Boltzmann and Gibbs is to have made just that sound plausibly *true*.

▲▲▲ And *that*, at last, is what seems to me to be more or less the right way of describing the overall structure of classical statistical mechanics. But there's still a bit more to be said. It happens that the actual history of reactions to the reversibility-objections is a good deal longer and more varied and more complicated than what we've just been through, and—notwithstanding the fact that we have already said how it is that the story rightly *ends*—there are some other parts of it, there are some detours and false starts, there are (more particularly) some attempts at defending the sort of probability-distribution we have just now gotten done rejecting, the probability-distribution which is uniform (on the standard measure) over the world's *present* macrocondition, which ought not to go altogether unmentioned here.

Here (to begin with) is an especially tortured paragraph from Gibbs, which I came across in a great book by Larry Sklar:

> But while the distinction of prior and subsequent events may be immaterial with respect to mathematical functions, it is quite otherwise with respect to events in the real world. It should not be forgotten, when our ensembles are chosen to illustrate the probabilities of events in the real world, that while probabilities of subsequent events may be often determined from probabilities of prior events, it is rarely the case

13. Or at any rate the entirety of that sector of the universe which has any physical interaction with the systems of interest to *us*, at the beginning of the epoch to which we have any epistemic *access*.

> that probabilities of prior events can be determined from those of subsequent events, for we are rarely justified in excluding from consideration the antecedent probability of the prior events.[14]

What's this supposed to *mean*? That the probability-distributions over microconditions are (as a matter of literal fact) nothing of the sort? That what they are (that the *entirety* of what they are) are serviceable *instruments* for the prediction of the *macro*future—that they have nothing directly to do with the *frequencies* with which *microconditions* are actually *realized in the world*? But if *that's* so, then there would seem to be a great deal of work still to be done. We are now going to be very curious to know exactly *how it is* that these instruments manage to *work*; we are going to be very curious to know (that is) *what* the actual frequencies of microconditions *are*. And of course both of those questions are quite neatly dealt with by *our own* latest version of the postulate about statistics—the one about the macrocondition of the world at its *outset*—but that can't have been the direction in which *Gibbs* intended to go!

Sklar thinks that some sense can be made of Gibbs here if we read him as espousing something along the lines of a purely *subjectivist* interpretation of the statistical-mechanical probabilities. He thinks Gibbs must be thinking of probabilities here as degrees of *belief*. He thinks Gibbs must be thinking that "although future events are not yet known to us, and hence have as their 'probability' only such probability as can be inferred by us from our statistical theorizing, past events, being known by direct evidence of their occurrence, are not subject to having probabilities attributed to them in this way." But it's hard to see how *that's* going to help. Given that the equations of motion we're talking about here are fully deterministic, a probability-distribution over possible present or past or future conditions—subjective or otherwise—is ipso facto a probability-distribution over conditions at all *other* times *as well*. And so if what Gibbs has in mind here is *any interpretation whatsoever* of the statistical-mechanical probability-distributions as probability-distributions over *possible present microconditions*, then the problem with the distribution that's uniform (on the standard measure) over the world's present macrocondition, or over what we *know* of the world's

14. Lawrence Sklar, *Physics and Chance* (Cambridge: Cambridge University Press, 1993), 199.

present macrocondition, is still that that distribution radically fails to match up with what we know of the *past*.[15]

Erwin Schrödinger had another angle. He urges us not to forget for even a moment that all that needs to be accounted for is that entropy never decreases as time flows forward. The thing to keep your eye on, the thing people haven't been careful enough about heretofore, the thing that's caused all the confusion (so says Schrödinger), is the *meaning* of the locution "time flows forward." Look (he says): it follows from the fully time-reversal-symmetric version of the fundamental insight of Boltzmann and Gibbs that if at some given instant two *separate* systems are both in non-maximal-entropy macroconditions, and if the probability-distribution over the microconditions compatible with those macroconditions is taken to be the standard uniform one, then the entropies of both systems are overwhelmingly likely to be going up toward the *future*, and the entropies of both systems are also overwhelmingly likely to be going up toward the *past*, which is to say, it is overwhelmingly likely that their entropies *always* evolve *in parallel* (as in Figure 4.3). And note that the two systems in question here can be *any* macrosystems *at all;* one of them (for example) can be taken to be the system of interest, while the other is the rest of the universe. And what (if you think about it) can "the direction in which time flows forward" really mean other than "the direction in which the entropy of the bulk of the universe, the direction in which the entropy of the natural *reference system*, is *increasing*." And so (says Schrödinger) there was never really any problem here to begin with.

Well, he must be kidding. There most certainly *is* a problem here, which is that the Boltzmann-Gibbs insight, together with the stipulation that the probability-distribution over the microconditions compatible with the world's present macrocondition is the standard uniform one, entails that what we seem to remember of the *past* is almost certainly *wrong*. And what Schrödinger has to say here does *nothing whatsoever* to alleviate *that*. What Schrödinger does here is (as it were) to eliminate the *problem* with the past by eliminating the past *itself*. But *that's* not what we need; what we

15. Sklar's *own* objection to Gibbs (or to his *reading* of Gibbs) is that in fact "past events are as frequently a matter of inference to us as are those of the future." But that seems utterly beside the point. Indeed, the more we *do* directly know about the past, the more there is going to be for the distribution that's uniform (on the standard measure) over the world's present macrocondition to fail radically to match up with, and the *worse* off Gibbs is going to be!

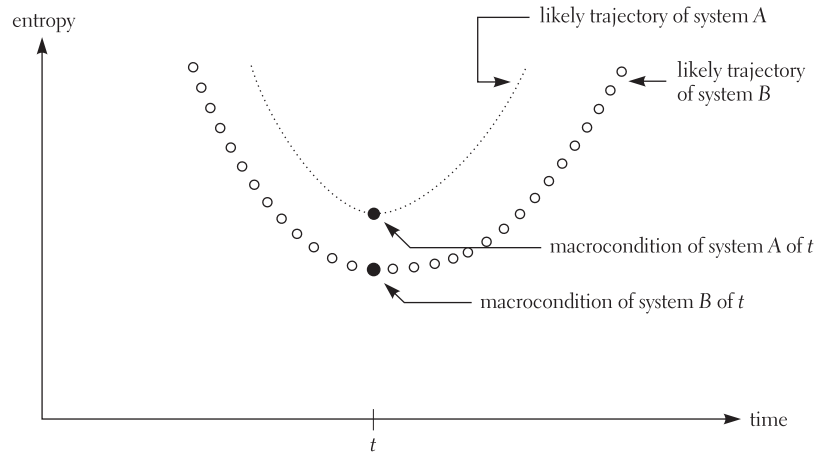


Figure 4.3

need, and what the solution I proposed above accomplishes, is to get the past *right*.

Paul Davies (who in *The Physics of Time Asymmetry* sees himself as taking up a train of thought that goes back to the great Hans Reichenbach) is a bit less off the mark. Davies and Reichenbach both understand very clearly that we will be getting some part of the past wildly wrong if the instant at which we choose to apply the uniform distribution over microstates is anything other than the one at which things *begin*. But *what* things, exactly? Davies seems to think that there are specifiable moments at which the quasi-isolated medium-sized systems of our *everyday thermodynamic experience*—the "branch" systems, as he calls them—can be said to begin. He seems to think that one can point to an exact moment in the history of any such system at which it was "formed by interaction with the outside world." He seems to think (for example) that a glass of water with a chunk of ice floating in it can be said to have *come into being* exactly at the moment when (I suppose) *the chunk of ice was dropped in*—and that the right probability-distribution over the possible microconditions of a system like that is (consequently) the one that's uniform (on the standard measure) over the region of that system's phase space which is compatible with its macrocondition *then*. And as to the past of *that* moment, well, the idea is clearly that "if a branch system is

formed in a random low-entropy state, it simply *did not exist*" in the past of that moment. And that's that.

But all this is sheer madness. How is it (to begin with) that we are to decide at *exactly* what moment it was that the glass of water with ice in it first came into being? And even if we *could* decide that, what then? How is it (exactly) that the medium-sized system we decided to focus on was the glass of water with the ice in it and not (say) the *room* in which that glass is currently *located*, which also contains the table on which the glass is currently sitting, and the freezer from which the ice was previously removed, and the person who first got it into his head to do the removing? The uniform probability-distribution over the possible microconditions of the macrocondition of *that* system, at the moment when *it* came into being, will (after all) differ quite radically—even insofar as the glass of ice water *itself* is concerned—from the one we have just been talking about! And why not the *building* that the room is *in*? And why not the *city* that the *building* is in? And even if all *that* could be decided, very serious questions would remain as to the *logical consistency* of all these statistical-hypotheses-applied-to-individual-branch-systems with *one another*, and with the earlier histories of the branch systems those branch systems branched off *from*. And all that aside, why in God's name *bother* with all this, when the uniform probability-distribution over the possible microconditions compatible with the macrocondition of the *world*, at the moment when *it* came into being, will very straightforwardly give us everything we need?[16]

▲▲▲ One more historical matter. The physical literature is positively infested with suggestions about what the "origin" of the lowness of the initial entropy of the world is, or about what it is that "drives" or "powers" or "explains" the world's constant entropy *increase*.

Thomas Gold, for example, is famous for noting that if the universe were

16. I think Davies must have something along the lines of an *epistemic* understanding of the statistical-mechanical probabilities somewhere in the back of his mind. That, at any rate, is what I read into locutions like "*random* low-entropy state" (emphasis mine). I think that sort of talk still has a ring of *innocence* for him. But if *that's* his game, if he wants to see himself as in some sense positing nothing at all here over and above the laws of Newtonian mechanics, then one must ask why this randomness should a priori come at the *beginnings* of the careers of these systems, rather than at their *ends*, say, or in their *middles*?

somehow confined to within a small spatial volume, then it would quickly reach equilibrium, and all temporal directionality would vanish. And this, of course, is true. And this (moreover) is a way of making clear in precisely what *sense* the initial state of the world was far from equilibrium—of which more in a minute. But Gold seems to see a hell of a lot more in it than that. He speaks of the expansion process (if I understand him) as what *drives* the entropy increase, as what *causes* it; as if that process could somehow *stand in* (I guess) for the *postulate about statistics*, as if the fact that the process is occurring could somehow bring about the entropy increase all by itself, *independent of the initial microconditions*. He seems to think (as a matter of fact) that the entropy of the universe would have no choice but to begin to *decrease* in the event that the universe should ever begin to *re-contract!* And this (it apparently needs to be said) is simply insane—there are plainly locations in the phase space of the world from which (on the Newtonian equations of motion, or on the general-relativistic equations of motion, or on any classical set of equations of motion you like) the world's radius will inexorably head *up* and the world's entropy will inexorably head *down*, and that's all there is to it, and there's nothing to do about it, period, end of story.

One reads (similarly) in authors like Paul Davies that "the *origin* of all thermodynamic irreversibility in the real universe ultimately depends on gravitation." And if what Professor Davies means is that the nature of those macroconditions whose entropies are particularly high or particularly low, in our world, is *much influenced* by the presence of gravitation, then he is surely right—and (moreover) he is making an interesting point. Consider (for example) how stars come into being. What happens (we think) is that an initially dispersed cloud of dust—under the influence of its own gravitation—*clumps up*. And it is as clear as it can be that the clumping up is a thermodynamically *irreversible* process. And so it must be the case (and here is something unfamiliar from our considerations of gasses—here is the influence of *gravity*) that the entropy of the clumped-up condition is *higher* than the entropy of the dispersed one, It must be (that is) that the dust particles are overwhelmingly likely to *pick up a great deal of momentum*, in all sorts of different directions, as they fall in toward one another—it must be (that is) that the clumped-up condition is the condition that's nonetheless by far the more *dispersed* one in the *mu*-space. But anything *more*, anything to the effect that gravitation can somehow *cause* entropy to increase more or less independ-

ently of the identity of the world's initial *microcondition*, anything to the effect that (say) the entropy of the world might suddenly start to *decrease* if gravitation were suddenly to turn *repulsive*, is (well) nuts.

Enough about all this. Let's get back on track.

▲▲▲ All the elements of a coherent classical statistical-mechanical account of the world are now (I think) on the table. And the only thing that still seems to me to need doing, by way of finishing this chapter up, is to raise a certain question—a question which (by the way) is going to reappear, in a slightly different form, as the central topic of Chapter 6—about the *epistemology* of that account.

It will take a little setting up.

Call the present macrocondition of the world *P*. And let M_P represent that region of the phase space of the world which is *associated* with *P*. And let D_P represent the probability-distribution which is uniform—on the standard measure—over the entirety of M_P, and is zero elsewhere. Now, D_P is of course (by construction, by *definition*) compatible with the present macrocondition of the world, and it is also (by virtue of the fundamental insight of Boltzmann and Gibbs) compatible with everything we have learned to expect of the *future*, and so the various *alterations* and *amendments* to D_P that we've been considering over the course of the present chapter have (once again) been aimed exclusively at bringing it into accord with what we take ourselves to know about the *past*. The trouble with D_P (remember) is that D_P and the laws of motion turn out to entail that you and I previously looked *older* than we do now, and that the building I am sitting in previously looked more *worn* than it does now, and that the half-melted ice cube in the glass of water in front of me was previously *more* melted than it is now, and (more generally) that the entropy of the universe was previously enormously *higher* than it is now—and the thing is that we are as certain as we are of anything in the world that that's all *wrong*.

And if it should ever occur to anybody to ask exactly where all that certainty *came from*, we would likely respond (at first) that the evidence is so abundant and so familiar and so unassailable that it's hard even to know where to start: we can show *photographs* (I guess) or play *tape-recordings* or point to *footprints* or dig up *fossil records* or look in *newspapers* or interrogate our *memories* or any number of other things like that—things as natural as

breathing, things we are all well accustomed to betting our very lives on, hundreds or thousands of times a day.

But there's something not altogether kosher about that.

Every last shred of the evidence we've been talking about is (after all) always already *part and parcel* of *P*. And so every last shred of this evidence is always already *automatically taken account of* in D_P. And when you think about *that*, it suddenly gets hard to see how this evidence can possibly amount to good grounds for any *alteration* or *amendment* of D_P *whatsoever*. And yet (and this is the rub) D_P is *precisely* the sort of distribution on which it is *overwhelmingly unlikely* that any of what we normally take this sort of evidence to *support* actually turns out to be *true!*

The thing is that D_P—together with the laws of motion—gives its *own* account of how the world is overwhelmingly likely to have gotten to be the way it currently (macroscopically) is, and *part and parcel* of that account is an account of how (say) this photograph I have "of myself at the age of 5" is overwhelmingly likely to have gotten to be the way it currently (macroscopically) is, and that account entails that the photograph was previously yellower and more worn than it is now, and that a very long time ago—with fantastic slowness—it formed, spontaneously, as a matter of pure chance, out of disparate wisps of paper and emulsion and dust, and that there has almost certainly never been anybody in the world who actually looked much like the boy in that picture.

The spontaneous formation of a photograph like that is of course an exceedingly *un*likely event on the probability-distribution which is uniform (on the standard measure) over the *unlimited entirety of the phase space of the universe*, or on the distribution which is uniform (on the standard measure) over the region of the phase space associated with the universe's having been in equilibrium a billion years ago, or on any distribution which is uniform (on the standard measure) over any region of the phase space corresponding to that photograph's *not having existed* (say) *fifty* years ago. But all that counts for absolutely nothing—on any of *those* distributions (after all) it is exceedingly unlikely that the present macrocondition of the world (the one with this photograph *in* it) actually obtains *at all!* The thing you have to start with—if you want to play by the rules—is what the macrocondition of the world currently, actually, *is*. And given *that*, and given that the probability-distribution over microconditions that we use in our calculations ought

to be the one that's uniform (on the standard measure) over the entirety of the region of phase space which is *compatible* with that present macrocondition, and given the insight of Boltzmann and Gibbs about the laws of motion—it is exceedingly likely, it is *overwhelmingly* likely, that the ice floating in the glass of water in front of me was more melted five minutes ago than it is now (notwithstanding my memory to the contrary), and that I looked older five years ago than I do now (notwithstanding the photograph I have in my desk), and that Napoleon never existed (notwithstanding what it says in the book in the next room). And I can of course produce as many *other* such photographs and books and eyewitness accounts as I like, and it will all be to no avail. It will be overwhelmingly likely that *all* of them came into being and ended up here in this room, together, by *coincidence*.

And so the flip-side of the insight of Boltzmann and Gibbs is that there can be nothing at all about the present macrocondition of the world which can possibly count as evidence that the world's entropy has ever previously been lower.[17]

And so the fact that the universe came into being in an enormously

17. Hans Reichenbach looks to me to be struggling rather desperately—on pages 129 and 130 of *The Direction of Time* (Berkeley: University of California Press, 1971)—not to believe this.

He seems to be thinking something like the following: consider (say) a spherical shell of light which is (at present) expanding outward from a star. And note that this expansion (unlike the melting of an ice cube floating in a glass of water, or the dispersing of smoke, or the aging of a human being) involves no increase in *entropy*. And so the *past* of a system like that—the past in which the shell is *smaller*—can reliably be inferred (unlike the past of a half-melted ice cube floating in a glass of water, or the past of a half-dispersed puff of smoke, or the past of a middle-aged human being) from its present macrocondition + the equations of motion + the standard statistical hypothesis. And so (Reichenbach imagines) the *image* of a younger, stronger, lower-entropy star which that light-shell *carries to our eyes* must be a reliable one too. And so it *can* reliably be inferred, after all, that the entropies of stars—and of galaxies, and of the world as a whole—were once indeed much lower than they are now.

And everything about all this is fine up until the last two sentences. The thing that seems to have slipped Reichenbach's mind is that although the process of the *expansion* of that sort of a shell of light into space is *indeed* reversible, the process of the *emission* of that shell from the star in the *first* place is emphatically *not*. What will follow about the past of that shell from its present macrocondition + the equations of motion + the standard statistical hypothesis is that (prior to the present) it was a *smaller* expanding shell than it is now, and that (sometime prior to *that*) it was all concentrated at a point which just happens to have coincided with the position (then) of the star in question, and that (sometime prior to *that*) it was a *contracting* shell, and that the fact that this shell has always carried with it an image of a young star is purely a matter of *coincidence*, and that it is in fact overwhelmingly likely that no such young star *ever existed*. And so it can certainly *not* be reliably inferred, in the way Reichenbach is thinking, that the entropy of the world was once much smaller.

low-entropy macrocondition cannot possibly be the sort of fact that we know, or ever *will* know, in the way we know of straightforward everyday particular *empirical* facts. We know it *differently*, then. Our grounds for believing it turn out to be more like our grounds for believing general theoretical *laws*. Our grounds (that is) are *inductive*; our grounds have to do with the fact that the proposition that the universe came into being in an enormously low-entropy macrocondition turns out to be enormously helpful in making an enormous variety of particular empirical *predictions*.

Suppose, for example, that we happen to dig up a decayed boot with an "N" embroidered on it. If the probability-distribution over microconditions that we use to make inferences about the world is the one that's uniform (on the standard measure) over those regions of the phase space of the universe which are compatible with everything we have thus far been able to directly observe of its present physical situation, then the probability we will associate with finding *another* such boot, if we dig around a bit further, will be overwhelmingly small. But if the distribution we use is the one that's uniform over those regions of the phase space of the universe which are compatible *both* with what we have thus far been able to observe of its present physical situation *and* with its having initially started out with a *big bang*, *then* that first boot will plausibly count as evidence for the truth of the proposition that there was once such a person as Napoleon, and the probability of our finding *another* such boot, if we dig around a bit further, will plausibly be much more substantial. And what our experience dictates is (of course) that the second of those predictions is much closer to the mark.[18]

Or think, again, about the case of spatulas. Suppose that I come upon an apartment about which I happen to have no direct empirical knowledge whatsoever other than the details of its architectural design and the fact that it contains a spatula. Just as before, if the probability-distribution over microconditions that I use to make inferences is the one that's uniform (on the standard measure) over those regions of the phase space of the universe which are compatible with everything I have yet been able to observe of its present physical situation, then the probability we will associate with finding

18. The previous two paragraphs (by the way) are very much in the spirit of the beautiful essay on the distinction between the past and the future by Richard Feynman in his book *The Character of Physical Law* (Cambridge: MIT Press, 1967).

the spatula in the bathtub comes out (as we saw in Chapter 3) too big. But if the distribution I use is the one that's uniform over those regions of the phase space of the universe which are compatible *both* with everything I have yet been able to observe of its present physical situation *and* with its having initially started out with a *big bang*, *then* (and *only* then) there is going to be good reason to believe that (for example) spatulas typically *get to be where they are* in apartments only by means of the intentional behaviors of *human agents*, and that what human agents typically *intend* vis-à-vis spatulas is that they should be in *kitchen drawers*.[19]

And note that we have just now more or less inadvertently done away with what (so far as I know) was the last existing impediment to a fully satisfactory formulation of the fundamental postulate about *statistics*. The postulate we *started out* with (remember) was the one according to which the correct probability-distribution over the possible microconditions of a system S, given that all I observationally know of S is X, and where X can be *any information whatsoever* about the present physical situation of S, is the one that's uniform (on the standard measure) over whatever region of the phase space of S that X is *compatible* with. And the trouble with *that* formulation (the trouble, that is, that became obvious on page 65) was that it generated propositions about the positions of spatulas which are *false*. And the best thing we could come up with by way of *straightening that out*, at the time, was to limit the Xs to *complete macrodescriptions*. And what that *left* us with (as we saw on page 67) was a postulate that *failed* to generate certain statements about the locations of spatulas—certain very *general* and *robust* and *law-like* statements about the locations of spatulas—which are *true*. And now (under the pressure of the altogether *different* sorts of considerations, considerations about the *past*, that have emerged in the *present* chapter) the postulate has changed yet again—to the effect that the probability-distribution that one ought to plug into the equations of motion in order to make inferences about the world is the one that's uniform, on the standard measure, over those regions of the phase space of the world which are compatible both

19. In the *absence* of any stipulation to the effect that the universe started out in a big bang—or in *some* very low-entropy state, at any rate—spatulas will typically come to be where they are in apartments by means of *random spontaneous materializations*, and the intentions of *human agents* vis-à-vis spatulas (supposing that it would make any sort of sense, under such circumstances, even to *speak* of such things) will typically be absolutely all over the place.

with *whatever it is* that we may happen to know about the present physical condition of the universe (just as in the *original* postulate) *and* with the hypothesis that the original macrocondition of the universe was the one associated with the *big bang*. And the thing that we have just now stumbled across is that this *third* formulation of the postulate appears to get the story about spatulas *just right*.

But of all this more later—our present task (which has just been to put all the elements of the classical statistical-mechanical picture of the world more or less in their proper conceptual boxes) is done.

It comes out like this:

The Newtonian statistical-mechanical contraption for making inferences about the world consists, in its entirety, of three laws and one contingent empirical fact.

The empirical fact is the one about what the macrocondition of the world currently happens to *be* (or rather, the empirical fact is the one about what the *directly surveyable* condition of the world currently happens to be; where the directly surveyable condition of the world—insofar as Mr. X is concerned—includes its macrocondition plus whatever, perhaps microscopic, features of Mr. X's *brain* he may happen to have direct and unproblematic introspective *access* to), and the laws are:

1. The Newtonian law of motion (which is that $F = ma$).
2. The *Past-Hypothesis* (which is that the world first came into being in whatever particular low-entropy highly condensed big-bang sort of macrocondition it is that the normal inferential procedures of cosmology will eventually present to us).
3. The *Statistical Postulate* (which is that the right probability-distribution to use for making inferences about the past and the future is the one that's uniform, on the standard measure, over those regions of phase space which are compatible with whatever other information—either in the form of *laws* or in the form of *contingent empirical facts*—we happen to have).

THE SCOPE OF THERMODYNAMICS

There was a question that came up at the beginning of Chapter 3—the question about whether the second law of thermodynamics had any systematic *exceptions*, the question (that is) about *Maxwell's demon*—which we are now in a position to think through with a good deal more care.

▲▲▲ Let's start a few steps back.

People have sometimes worried that (say) the development of a human infant from a single fertilized cell might somehow amount to a violation of the second law of thermodynamics. The worry is obviously that processes like that seem to correspond to decreases in entropy; the worry is that the *ends* of processes like that seem to represent fantastically more *organized* conditions of the organic material involved than their *beginnings* do. And it is by now a cliché of the thermodynamical literature that this sort of worry is utterly *misguided*; that this sort of worry can only come up if you have been thinking of the organic material in question as *isolated*. And it *isn't* isolated, of course, and it isn't a particularly difficult business to argue that as a matter of fact the total entropy of a developing fetus and its mother and her environment goes radically *upward* in the course of that development, and so the whole business is no more mysterious, and no different in kind, than (say) the decrease in the entropy of a hot body which is put in thermal contact with a cooler one. And the same thing goes for the thermodynamic status of biological evolution as a whole, and for the thermodynamic status of the construction of cities, and for the thermodynamic status of the electronic organization of information, and so on.

Good.

Here's another sort of worry, which gets undone (in the end) by means of

a particularly beautiful observation, which we owe (insofar as I know) to Richard Feynman.

Consider a contraption like the one pictured in Figure 5.1. There is (to begin with) a box with some gas in it. And there is an axle. And one end of the axle is sticking into the box, and that end has some plates or sails or vanes (or whatever you call them) attached to it. And the other end is fixed up with a spring-loaded ratchet-and-pawl sort of a thing that's designed to allow the axle to turn only in a single direction—clockwise, say. And the middle is attached to a string with a little weight on the bottom.

And the way the contraption is intended to work is this: the gas molecules will for the most part be banging more or less equally on each of the two sides of each of the vanes, and (consequently) pushing them nowhere. But every now and then, by pure chance, that won't be so. Every now and then (that is) there will just happen to be more gas molecules banging the vanes in (say) the *clockwise* direction than in the *counterclockwise* direction—and the probability of there *being* an imbalance of that sort, of a certain particular *degree*, per unit *time*, is clearly going to go up with the gas's *temperature*. Anyway, when such an imbalance occurs, the axle will *turn*, and the weight will *lift*, and the temperature of the gas will *decrease*; and if we wait long enough the axle will almost certainly turn *again*, and lift the weight *higher*, and the temperature of the gas will decrease still *more*. And of course the ratchet-and-pawl mechanism will guarantee that none of that turning or lifting or cooling ever gets *undone*; and so what we apparently have here is a means for extracting any amount of heat we like from a gas which is at uniform temperature throughout, and converting it entirely into mechanical energy, and (in the course of all that) bringing about no other thermodynamic changes whatsoever in the world, which is directly in violation of Kelvin's formulation of the second law.

But look a bit more closely.

Note, to begin with, that if the ratchet and pawl were to be manufactured (as it were) *too well*, if the ratchet and pawl were to be manufactured in such a way (of which more in a minute) as to guarantee that their combined total mechanical energy never *decreases*, then whenever the pawl happens to snap to the bottom of a new ratchet-tooth it will have no choice whatsoever—by simple conservation of energy—but promptly to bounce *all the way back up*

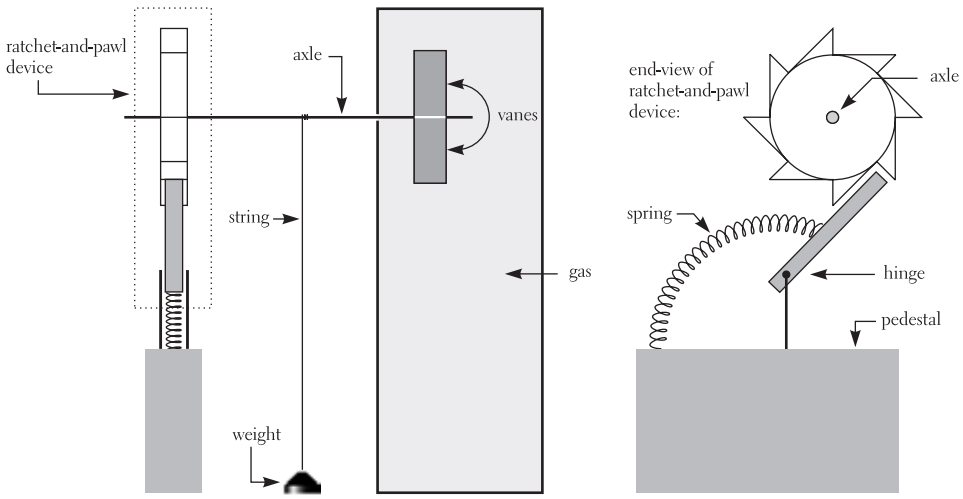


Figure 5.1

again. And then (of course) the weight will promptly pull the ratchet *back*. And then the whole contraption will be useless.

And so (if this contraption is going to work) the normal operational interactions between the ratchet and the pawl are going to have to involve some sort of *rubbing* or *scraping* or *denting* or something like that—they are going to have to involve the systematic transformation of some of the mechanical energy of the ratchet-and-pawl system into *heat*. And so the temperature of the ratchet and pawl, in the course of each turn-step, is going to have to *rise*. And so the *entropy* of the ratchet-and-pawl system, in the course of each turn-step, is going to have to rise.

And remember (from Chapter 2) that the *amount* by which the entropy of any macroscopic system rises in the course of any thermodynamic transformation is (by definition) equal to the amount of heat it absorbs in the course of that transformation *divided* by its *temperature* at the *time* of that transformation. And note (in that connection) that the proper functioning of this contraption is going to require that the *temperature* of the ratchet and pawl *not be too high*. The higher the temperature—you see—the higher the probability per unit time of statistical fluctuations whereby (say) the pawl sponta-

neously lifts *itself* up off the ratchet, and so leaves the ratchet free, for a moment, to turn backward, and so (again) renders the contraption useless.

And what turns out to *follow* from all that—without too much more trouble—is that the amount by which the entropy of the ratchet and pawl is going to have to rise in the course of each turn-step will necessarily *exceed* the amount by which the entropy of the *gas* is simultaneously going *down*.

And so the second law is saved again.

▲▲▲ Now to the main event.

Here's the setup: there's an isolated system called S that consists of two gasses, and of the box those two gasses are in, and of the interior wall between them, and of the movable shutter in that wall (all of which is pictured in Figure 3.3) and of a very talented but thoroughly physical "demon." And the temperatures of the two gasses are initially different. And the demon is ready and willing and able and implacably disposed to carry out a carefully coordinated program of measurements and calculations and manipulations of the shutter—the sort of program (that is) that we discussed at the beginning of Chapter 3—which is designed to make those temperatures differ still *more*.[1]

And the problem (as mentioned in Chapter 3) is just that this sort of program looks very much as if it ought to *work*. The problem (more particularly) is that under the circumstances described above, what the three fundamental laws of statistical mechanics look very much as if they will entail[2] is that the entropy of S is overwhelmingly likely to go *down* toward the future.[3] And that—once again—looks to be in direct and flagrant violation of the so-called second law of thermodynamics.

And once again there is a famous collection of arguments in the literature to the effect that all these "looks" are (as a matter of fact) *illusory*.

The idea at the *center* of those arguments (which is exactly the one criticized at the end of Chapter 3, but of that more later) is that entropy is an

1. Suppose (that is) that he is initially disposed to measure and to calculate and to manipulate the shutter so as to allow only very high velocity molecules to pass from the cooler gas to the warmer one and so as to allow only very low velocity molecules to pass from the warmer gas to the cooler one.

2. That is, the three fundamental laws of the *world*, the ones I wrote down—after many false starts—at the very end of Chapter 4.

3. As a matter of fact, laws 1 and 3 *alone* will apparently entail that; and law 2 will simply do nothing to *contradict* it.

epistemic business, that entropy is a characteristic of *probability-distributions*, that the entropy of any given system at any given time is (roughly speaking) a measure of the number of distinct microconditions which—for all we happen to *know*—that system at that time might imaginably be *in*. And it happens that if the evolution of any system *S* over the course of any interval *I* is isolated and unobserved, and if the laws that govern *S*'s evolution are deterministic in *I*, and if the laws that govern *S*'s evolution are symmetric under time-reversal in *I*, then it follows (about which more in a minute)—absolutely irrespective of any further details about precisely *what* sort of a system *S* happens to *be*—that the size of the set of microconditions which *S* might imaginably be in can simply not *decrease* in *I*. And that's more or less all there is to it.

Take the particular case of the demon. If everything goes as advertised, the demon is supposed to be able to accept any two-gas system we happen to hand it in some particular macrocondition *A*—whose entropy is relatively high—and to hand it back to us sometime thereafter in some *other* particular macrocondition *B*, whose entropy is *lower*. If everything goes as advertised (to put it another way), and if all we initially *know* of the two gasses in question is that their joint macrocondition is *A*, then what the demon is supposed to be able to do is (roughly speaking) to *reduce* the number of distinct microconditions that the two-gas system might imaginably be *in*. If everything goes as advertised, then (speaking roughly again) there are supposed to be at least two distinct microconditions of the two-gas system—*c* and *d*—both of which are compatible with *A*, and *either one* of which the demon is able to transform into some unique *third* microcondition, *e*, which is compatible with *B*.

And it turns out (and this is the punch line) that every such *bringing together* of distinct conditions in the two-gas system can only be accomplished at the cost of a corresponding bringing *apart* in the demon *himself*. Think about it: the particular things this demon is going to need to do in order to transform *c* into *e* are necessarily going to be *different* from the particular things he is going to need to do in order to transform *d* into *e*.[4] And so the robot-demon is going to need to ascertain and store in his memory some-

4. Note, by the way, that it is absolutely crucial to the truth of this claim that the final states in these two transformations are *the same*. It might very well be the case, in principle, that the things that need to be done in order to transform *c* into *e* are *precisely the same* as the things that need to be done in order to transform *d* into (say) *f*.

thing about *which particular one* of those two initial conditions actually *obtains* in order to decide how to *proceed*; he is going to need to ascertain and store in his memory something about (say) the positions and velocities of the particles in the vicinity of his shutter—in the case of the scheme Maxwell discussed—in order to know precisely when that shutter will need to be *opened* and *closed*. And so (if you think it over) for every two distinct conditions compatible with *A* which the robot-demon is able to bring together in *B*, there are necessarily going to be two distinct conditions in which the demon's own physical *memory-elements* may potentially *end up*. And so as the number of distinct conditions that the two-gas system might imaginably be in decreases, the number of distinct conditions that the robot-demon *himself* might potentially be in will necessarily be going *up*.[5] And those numbers (once again) are precisely the ones that the entropies of those systems are *about*. Or so the argument goes.

Let's sharpen all this up a bit. Strictly speaking, the talk in the last several paragraphs about the "number" of distinct microconditions that a certain system might imaginably be in, or about the "number" of distinct microconditions compatible with this or that macrocondition, is all wrong—the number of such distinct microconditions (after all) is invariably infinite. The *right* thing to talk about in this connection—if you want to be careful—isn't *number*, but *measure*. And the particular measure on sets of microconditions that thermodynamic entropy happens to be connected *with*—as we saw in Chapter 2—is the standardly calculated *volume* in *phase space*. And so the thing that apparently needs to be demonstrated in order to establish that the entropy of *S* cannot decrease in the course of the exercise we've been talking about (*whatever* might or might not happen, in the course of that exercise, to the two-gas *subsystem* of *S*) is just that the standardly calculated phase-space volume of the set of distinct microconditions that *S* might imaginably be in is no smaller at the *conclusion* of the exercise than it is at its *beginning*.

5. And it hardly needs saying (and this has nonetheless recently been made a very big deal of in the physical literature) that if there happens to be some *other* mechanism in the picture, whereby the demon can be *reset*, when his work is done, to his *original* microcondition, then the number of distinct microconditions that the *resetting mechanism* might potentially be in will necessarily go *up*—in the course of that resetting—by precisely the same amount as the number of distinct microconditions that the *robot-demon* might potentially be in goes *down*. And so on and so on (if there should also happen to be resetting mechanisms for the resetting mechanisms) ad infinitum.

And it happens to be *exactly the content of the Liouville theorem* (you will remember) that the standardly calculated volume in phase space of any set of microconditions of any isolated system will be exactly equal to the standardly calculated volume in phase space of the set of whatever *other* microconditions those original ones get *carried into*, over any particular time-interval, by the Newtonian equations of motion. And so—if entropy is indeed the sort of thing that we've been *taking* it to be over the last few paragraphs—then the entropy of an isolated and unobserved Newtonian-mechanical system can patently *never go down*. Period. End of story. But the rub (now) is that it can just as patently also never go *up*. And it *does* go up, of course. And so something is terribly wrong.

▲▲▲ Let's start again.

The thing, of course, is that entropy is *not* an epistemic business. Entropy (as I went to some trouble to point out at the end of Chapter 3) is an objective physical characteristic of the individual microconditions of individual thermodynamic systems. The entropy of a microcondition is the logarithm of the standardly calculated phase-space volume of the macrocondition to which the microcondition in question *belongs*. And if (as often happens, in our everyday experience) the microcondition of a certain thermodynamic system happens to wander from a *smaller*-volume macrocondition into a *larger*-volume macrocondition, then (at that instant) the entropy of that system goes *up;* and if (as *rarely* happens, in our everyday experience) the microcondition of a certain thermodynamic system happens to wander from a *larger*-volume macrocondition into a *smaller*-volume macrocondition, then (at that instant) the entropy of that system goes *down*. Period.

And once that sinks in, and if (as usual) we take the fundamental microscopic dynamical laws of the world to be the Newtonian ones, then it follows almost immediately that the proposition that the entropies of isolated thermodynamic systems do not decrease *cannot be a universal law;* not a strict one (which is old news, of course), and not a *statistical* one, *either*.

Imagine (for example) a certain demon which is capable of surveying and (thereafter) rearranging the microconditions of isolated boxes of gas in such a way that at a certain particular time *after those rearrangements have been completed,* the entropy of the gas will begin spontaneously to *decrease*. Imagine (more precisely) that there is a demon with at least one macrocondition

M and a gas with at least one macrocondition *A* such that if at some initial time the demon is in *M* and the gas is in *A*, and the demon and the gas are in the appropriate sort of proximity to each other, then it is overwhelmingly likely that at a certain later time—a time after the demonic rearrangements are complete—the entropy of the gas (then isolated!) will begin to go down. Imagine (that is) a system *D* with at least one macrocondition *M* such that almost the entirety of phase-space volume of the macrocondition

{*D* is in *M* and *G* is in *A* and *D* and *G* are in the appropriate sort of proximity to each other}

is taken up with microconditions which the deterministic Newtonian equations of motion entail will begin to move, once the time-interval in question has elapsed, once (that is) *G* is evolving in complete isolation, through regions of the phase space associated with progressively lower *entropies* for *G*.[6]

And note that the compatibility of the existence of a demon like that with the Newtonian equations of motion is an absolutely uncontroversial matter—all that the controversy was ever about (remember) is the amount by which the entropy of a demon like that would need to *go up* in the course of its operations. Let it need to go up (then) by any amount you like. It doesn't *matter*. The game (in a certain technical sense, at any rate) is already *over*. If such a demon can exist, and completely irrespective of the amount by which the entropy of a demon like that would need to increase in the course of its operations, the second law, in all its various formulations, turns out not to be universally true. If such a demon can exist—which nobody denies—then it cannot be a matter of anything like a universal law of nature that the entropy of an isolated gas which is not in equilibrium with respect to its gross constraints is bound or even statistically likely to go up over the next few minutes.

6. The "overwhelmingly likely" and "almost the entirety" locutions are (by the way) absolutely crucial here. Nobody (after all) disputes the *physical possibility* of the entropy of an isolated thermodynamic system's going down; what we're after here are circumstances under which the decrease of the entropy of an isolated thermodynamic system is not merely *possible*, but precisely what one ought to *expect*. And the same sorts of considerations apply to the insistence that *M* be a *macrocondition* (as opposed to a *microcondition*) of *D*: if all this is supposed to *work*, if all this is supposed to demolish the generality of the truth of the second law of thermodynamics, then *M* had better turn out to be the sort of thing that we can actually *prepare*, the sort of thing to which the laws of thermodynamics are actually meant to *apply*; it had better turn out (that is) to be a *macrostate*.

Whether that will happen will depend (rather) on whether there happen to have been any such demons *around* at certain particular times in the *past!*

These (however) are plainly not quite the sorts of demons that either Maxwell or his critics had in mind. The demons we've just been talking about (after all)—notwithstanding that they are statistically reliable producers of violations of the literal prohibitions of the second law of thermodynamics—will not necessarily leave the entropy of the world any lower (once the whole business is over) than it was back when we first decided to put them to work.[7] And so they aren't the sorts of demons that can make you any *money*—they aren't the sorts of demons (that is) with whose help one can reliably increase that portion of the total energy of the world which is available for our exploitation by gross mechanical procedures. Let's call then *pseudo*--Maxwellian demons, then.

▲▲▲ And now (at last) let's get to the real thing. Call him (as above) *D*. And call the system consisting of the two gasses and the box and the dividing wall and the shutter *G*. *D* + *G*, then, is the system we were referring to above as *S*. And now the question about whether or not there can *be* a demon of the sort that Maxwell was imagining comes down to something like this: can there be a system *D* with at least one macrocondition *M* such that if *D* is in *M* and *G* is in *A* and if *D* and *G* are brought into the appropriate sort of proximity to each other, and are subsequently left alone, then the macrocondition of *G* at a certain particular later time is overwhelmingly likely to be *B, and* the entropy of *D* at that particular later time is overwhelmingly likely *not to have gone up*. Can there be a system *D* with at least one macrocondition *M* such that (in other words) almost the entirety of phase-space volume of the macrocondition

{*D* is in *M* and *G* is in *A* and *D* and *G* are in the appropriate sort of proximity to each other}

is taken up with microconditions which the deterministic Newtonian equations of motion entail will move, over the time-interval in question, into re-

7. This (once again) is because we have put no restrictions whatsoever here on how much the entropy of the demon *itself* may need to *rise* in the course of its operations.

gions of the phase space associated with G's being in *B and with D's being in some macrocondition whose entropy is not higher than the entropy of M?*[8]

And at first glance, the answer would still appear to be *no*.

The sort of thing that occurs to you is this: the volumes in phase space of any particular initial set of microconditions of any isolated system *S* and of any particular one of the *time-evolutions* of that initial set will necessarily—by Liouville's theorem—be equal.[9] And remember (and this is just a trivial characteristic of the geometry of phase space) that for any thermodynamic systems S_1 and S_2 and any macrocondition $\{c_1\}$ of S_1 and any macrocondition $\{c_2\}$ of S_2, the volume of the macrocondition $\{S_1$ is in c_1 and S_2 is in $c_2\}$ in the phase space of (S_1 and S_2) is just the volume of the macrocondition $\{c_1\}$ in the phase space of S_1 multiplied by the volume of the macrocondition $\{c_2\}$ in the phase space of S_2. And so there can certainly *not* be a system *D* with a macrocondition *M* such that the overwhelming majority of phase-space volume of the above-mentioned macrocondition is taken up with microconditions which will move, over the time-interval in question, into regions of the phase space associated with *G*'s being in *B* and with *D*'s being in some particular macrocondition whose entropy is not higher than the entropy of *M*.

But (notwithstanding superficial appearances to the contrary) that doesn't quite settle the question. Let's keep our eye on the ball. What we want to know (remember) is whether there can be a contraption which is capable of reliably reducing the entropy of an isolated system of which that contraption itself forms a part; what we want to know (more particularly) is whether there can be a contraption capable of reliably taking the two-gas system we've been talking about from A to B—as Maxwell envisioned—without increasing its *own* entropy in the process. And note (and this is why the conclusion of the previous paragraph is utterly beside the point) that nothing in any of this requires that the final macrocondition of the demon be *the same on every run*. Nothing in any of this (that is) requires that the final macrocondition of the demon not depend on the details of the initial *microcondition* of the *two gas-*

8. *A* and *B* are of course just the (respectively) higher-entropy and lower-entropy macroconditions of the two-gas system we were talking about a few paragraphs ago. And note that all the stuff in footnote 6 about measure and likelihood and macroscopicness plainly needs to be borne in mind in connection with the present considerations as well.

9. And what I mean, of course, by "any particular one" of the time-evolutions is just the *n-second* evolution for any real value of *n*.

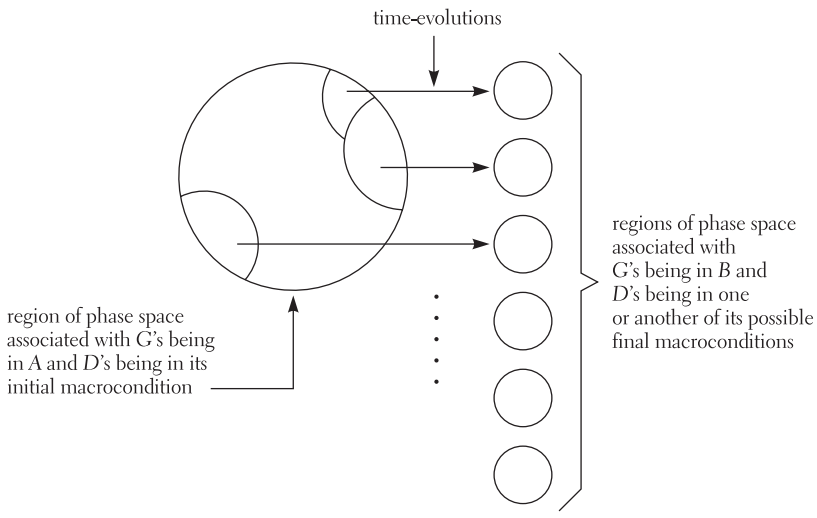


Figure 5.2

ses. And if the cardinality of the set of possible final macroconditions of the demon is *greater than one*, then (as Figure 5.2 makes clear) nothing whatsoever stands in the way of there being a macrocondition *M* such that the almost the entirety of phase-space volume of the above-mentioned macrocondition is taken up with microconditions which will move, over the time-interval in question, into regions of the phase space associated with *G*'s being in *B* and with *D*'s being in *one or another* of some *set* of macroconditions, *each* of whose entropies is no higher than the entropy of *M*.

What Liouville's theorem ineluctably requires is that the decrease in the volume of the region of phase space associated with the macrocondition of the two-gas system over the course of this exercise be *paid* for. But what has perennially been overlooked (it seems to me) is that there are *any number* of forms that the payment can legitimately *take*; and nothing whatsoever requires (in particular) that it take the form of an increase in the volume associated with the *macrocondition of D*.

Think (for example) of a Maxwellian demon whose individual physical memory-elements happen to be solid movable macroscopic bodies. Think (more particularly) of a demon who records the outcomes of his measurements of the positions and velocities of certain of the particles in *G* (so as to

be able to open and close his shutter at the appropriate times) exclusively in the macroscopic positions of a gigantic array of *billiard balls*. The entropy of a demon like that (and as a matter of fact, the *full thermodynamic condition* of a demon like that) could in principle *not change at all* in the course of his getting his job done.[10]

And so there are physical systems whose existence is perfectly compatible with the three fundamental microphysical laws of the world discussed at the end of Chapter 4—which (on the Newtonian picture) are all the strict or statistical fundamental laws there *are*—which have what you might call a *robust capacity* (which is to say, a capacity which is every bit as independent of the microscopic details of initial conditions as any *thermodynamic* behaviors are) *to lower the entropies of certain larger isolated systems of which they form parts*. There can perfectly well (in other words) be such things, such *physical* things, as bona fide *Maxwellian demons*.

And we have learned something interesting (by the way) about what *sorts*

10. Suppose somebody were to object, at this point, that it is part and parcel of *what it is* to be a macrocondition, that it is part and parcel of what it is (that is) to be the sort of condition to which thermodynamic predicates—predicates of entropy, for example—can even be coherently *applied*, that the number of distinct macroconditions that a system can potentially be in is relatively *small*. It seems to me that I've heard people say that sort of thing. And (you see) the number of distinct potential macroconditions that we are going to need to attribute to this set of *billiard balls*—if everything is to work as advertised—is more or less on the order of the number of distinct potential *microconditions* which are compatible with the initial macrocondition of the two *gasses*. And that *latter* number is (of course) *enormous*. And so (the objection will run) predicates of entropy simply *do not apply* to systems like the demon we've been thinking about here, and so they don't apply to systems like S (of which that sort of a demon is a subsystem) either, and so propositions to the effect that a system like S either obeys or fails to obey the second law of thermodynamics are simply *devoid of meaning*.

The thing is that any attempt at taking an objection like this *seriously* is going to leave us with a rather badly withered-up version of the second law; a version (that is) which is *enormously* more restricted and more uninformative and more uninteresting than the one we're used to thinking about; a version which (for example) is going to have nothing whatsoever to say about the possibility that somebody is eventually going to be able to slap together a contraption which is reliably capable of transferring heat from a hotter body to a cooler one without any attendant cost in mechanical energy to the rest of the world.

And anyway, and more important, and more trivially, the objection is *absurd*. On any reasonable construal whatsoever of what the word "macroscopic" is supposed to mean, any billiard ball in any infinite Newtonian universe can damn well potentially be in an *infinite* number of macroscopically distinct *locations*. And if entropy is an additive quantity (and it *is* an additive quantity), and if every one of the billiard balls of which the sort of demon we've been talking about is composed has a perfectly well defined entropy (and every one of them *does* have a perfectly well defined entropy), then—as a matter of *logic*—the entire finite collection of them damn well has a perfectly well defined entropy too!

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.[11] And that's what's *different* about the case of the demon, and that's why there turned out to be no genuine possibility of reducing the total entropy of the world by means of (say) the ratchet-and-pawl mechanism I was talking about before: the thing is precisely that the demon is *sensitive*, and the ratchet and pawl *isn't*.[12]

11. Note that all this is in no way incompatible with the sort of robustness discussed in the previous paragraph; indeed, what I am asserting here is precisely that microscopic sensitivity is a *necessary condition* of that robustness.

12. There can be no such thing (then) as a mechanism which reliably reduces the entropy of an otherwise isolated pair of gasses and that also reliably ends up in the macrocondition it *starts off* in, there can be no such thing as a mechanism which reliably reduces the entropy of an otherwise isolated pair of gasses and that also reliably ends up in precisely the macrocondition (that is) that constitutes its being all set to *start again*, on *another* such pair of gasses. And it has sometimes been thought, or at any rate it sometimes *seems* to have been thought, that a mechanism which reliably reduces the entropy of an otherwise isolated pair of gasses, and which suffers no increase of its own entropy in the process, *but which fails to end up in the macrocondition it started off in*, is somehow not quite legitimately or genuinely or fundamentally in violation of the second law of thermodynamics. And if anything like *that* were the case, then of course everything I've been talking about for the past several pages would be altogether *beside the point*. It isn't, but it will be worth another couple of sentences (I think) to inquire a bit more carefully into what a thought like that can possibly have been *about*, and how a thought like that can possibly have *arisen* in the *first* place.

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 system happens to return to the macrocondition it *starts out* in or not—period, end of story. The puzzle (as I said above) is about how anybody can ever have thought otherwise. And insofar as I can tell, the essence of the confusion has been a certain perennial equivocation in the literature about the meaning of the word "cycle."

Now and then—and particularly in the more old-fashioned discussions—the word "cycle" appears in formulations of the second law. Sometimes (for example) one sees Clausius' version of that law formulated as a proclamation of the impossibility of the existence of any physical contraption for transferring a given quantity of heat from a hotter body to a cooler one which (moreover) *operates in a cycle*. And what it is—in the present context—for such a contraption to "operate in a cycle" is for its *thermodynamic* parameters, and nothing whatever *other* than its thermodynamic parameters, to have the same values at the *end* of the heat transfer in question as they do at its *beginning*. And (as a matter of fact, and as a quick perusal of the Appendix will confirm) is turns out to be absolutely crucial to the standard demonstrations of the *entropic* version of the second law, it turns out to be absolutely crucial (that is) to the arguments from the

▲▲▲ A few concluding remarks.

First, the way the phase space of the world gets carved up into macroconditions depends very much (as mentioned in Chapter 3) on *us*; it is a matter of our *practices*, it is a matter of what we find we are easily able to *distinguish*, it is a matter of what we find we are easily able to *prepare*. And so the *entropies* of particular *points* in that phase space (which are just the logarithms of the volumes of the macroconditions to which those points belong) depend on us too. And yet (and this is the punch line) the above conclusions about Maxwellian demons do *not*. Here's the deal: you tell me how you want to carve the phase space up into macroconditions and I will be able to come back to you with a design for a Maxwellian demon which will be robustly capable of lowering the entropy (as *you* calculate it) of a larger isolated system of which it forms a part. If billiard balls won't do on the way you carve things up, we can just design a demon who records his memories in the positions of buildings, or cities, or planets, or whatever.

Second, the entirety of the foregoing discussion more or less straightforwardly carries over to relativistic theories and to quantum-mechanical theories and to relativistic quantum field theories and to relativistic quantum *string* theories and (in short) to every one of the theories that anybody has taken seriously over the past two hundred years or so as a candidate for the fundamental physical theory of the world. And among those theories the case of quantum mechanics deserves special mention, since there have been

---

Clausius and Kelvin formulations of the second law to the conclusion that the entropy of a system is a function of its present thermodynamic state *alone*, and not of how that state happens to have *arisen*, that nothing whatever *other* than the purely thermodynamic parameters of the contraption in question can come into play here! And the thing is that all this has somehow again and again been understood *in the context of discussions of the supposed impossibility of Maxwellian demons* in an altogether different way. The thing (more particularly) is that the requirement that the contraption in question "operate in a cycle" has somehow again and again been understood in the context of discussions of Maxwellian demons not merely as requiring the contraption's *thermodynamic* parameters return, at the end of the transformation, to their original values, but (rather) as imposing the enormously *stronger* constraint that the contraption return at the end of the transformation to its *full original macrocondition*, that it be ready, at the end, to start *over!* But (for the last time) that understanding is *wrong*. On that understanding (for example) the deduction of the entropic formulation of the second law from its Clausius and Kelvin formulations would simply not *go through*. On that understanding (to put it another way) the second law would not preclude (either strictly or as a statistical matter) decreases in the entropies of isolated thermodynamic systems. And precluding just that, of course, is what the second law is all about!

suggestions around more or less from the beginnings of recorded history that the quantum-mechanical uncertainty-relations might somehow make it impossible—as a matter of fundamental principle—for any such demon to carry out the sorts of *measurements* he needs to, or to operate his *shutter* in the way that he needs to. And there isn't a hell of a lot to say about those suggestions—without going into a good deal more detail than would turn out be either useful or entertaining here—other than that they are categorically *false*. It is certainly true (mind you) that in the quantum-mechanical case the question of precisely *what measurements the demon will need to perform* and the question of precisely *how his shutter will need to work* need to be thought through with a bit more *care*. But it isn't as if that's not a *perfectly accomplishable thing*—indeed, there has been an explicit *answer* to those questions, there has been (that is) an explicit quantum-mechanical *model* of those measurements and that shutter, sitting in John von Neumann's seminal book *Mathematical Foundations of Quantum Mechanics*, for anybody who might have cared to look, ever since the 1930s.[13]

Third, it deserves to be emphasized that there is nothing at all abstract or restricted or esoteric or legalistic or otherwise suspicious about the sorts of violations of the second law that these bona fide Maxwellian demons (as opposed to the *pseudo*-Maxwellian demons we were discussing before) are capable of producing. It is of the very essence of the thermodynamical definition of entropy (after all) that any decrease whatsoever in the entropy of any thermodynamic system invariably increases the amount of mechanical energy which can be extracted from that system by means of a heat engine, and (moreover) that any *increase* in the entropy of any thermodynamic system invariably *decreases* the amount of mechanical energy which can be extracted from that system by means of a heat engine. And so any reliable method of decreasing the entropy of one thermodynamic system without in the process *increasing* the entropies of any *others* (which is precisely the sort of thing that we have just now learned that bona fide Maxwellian demons can do) is nec-

13. Notwithstanding all this (by the way), von Neumann *himself* was convinced—by the sorts of arguments dispensed with above, the arguments based on an *epistemic* conception of entropy—that Maxwellian demons were physically impossible. John von Neumann, *Mathematical Foundations of Quantum Mechanics*, trans. R. T. Beyer (Princeton: Princeton University Press, 1955). (The first German edition was published in 1932.)

essarily a reliable method of increasing that part of the total energy of the universe which is available for mechanical exploitation. Period.

▲▲▲ The thing now is (obviously) to look for a catch. And the one thing that jumps right out at you in that connection is this issue of *sensitivity*, this fact that there can be no such thing as a Maxwellian demon that reliably *ends up* in *any particular macrocondition*, that (more particularly) there can be no such thing as a Maxwellian demon that reliably ends up in the macrocondition it *starts off* in, that there can be no such thing as a Maxwellian demon that reliably ends up in precisely the macrocondition that constitutes its being all set to *start again*, that there can be no such thing as a Maxwellian demon that operates—in the *stronger* of the two senses I described in footnote 12—*in a "cycle."* And the thought is that it might be possible to parlay that fact (together with the theory of relativity, perhaps, or with the laws of quantum mechanics, or with sciences of complexity, or with all of them in conjunction, or with something else altogether) into a demonstration that the construction of a Maxwellian demon system, or the *operation* of a Maxwellian demon system, or the *extraction of mechanical energy* from a Maxwellian demon system by means of heat engines (once its operations are done), or the *exploitation* of that energy (once it's been extracted) will necessarily—as a statistical matter—somehow prove prohibitive or self-defeating or otherwise uncircumventably pointless. The thought (to put it another way) is that there might be some interesting *generalization* of the second law which is true not merely in the absence of the sorts of circumstances we've been discussing here, but (rather) *simpliciter*; that there might be some interesting generalization of the notion of *entropy*, which (say) would reduce to the *familiar* notion in those circumstances in which the entropy is in fact statistically unlikely to go *down*, and which (as a matter of absolutely universal law) is *always* statistically unlikely to go down *itself*.

And insofar as I am aware, nothing whatsoever is currently known as to whether or how any of this can actually be made to *pan out*. And it seems to me a rather profound and urgent and fascinating question.

THE ASYMMETRIES OF KNOWLEDGE AND INTERVENTION

1. KNOWLEDGE

The sort of epistemic access we have to the past is different from the sort of epistemic access we have to the future. This (to put it mildly) nobody doubts. And nonetheless (and this is the first matter I want to take up in this chapter), there is a vast physical and philosophical literature nowadays about the alleged difficulty of specifying exactly *what that difference is*.

It's often pointed out, for example, that the difference certainly does *not* consist in our having knowledge of the past but none of the future. We *do*, after all, have knowledge of the future. We know (for example, and not less certainly than we know much of what we know of the past) that the sun will rise tomorrow.

And if it's said that we know *more* of the past than we do of the future, this seems (according to the usual way of talking) true enough, but (as it stands) not particularly informative—it seems to give us nothing at all that we can *reason any further* with, nothing that (as it were) we can *sink our teeth* into.

Sometimes the focus is shifted to differences between the *methods* by which we *come* to know things about the past and the future. It's said (more particularly) that there can be such things as *records* only of the past; but this is almost always immediately followed up with whining about the perennial elusiveness of exactly what it means to *be* a "record,"[1] and cluelessness follows again.

It seems to me (though) that things aren't nearly as mysterious as all that.

1. Have a look, for example, at what Larry Sklar has to say about it beginning on the bottom of page 385 of *Physics and Chance: Philosophical Issues in the Foundations of Statistical Mechanics* (Cambridge: Cambridge University Press, 1993).

The briefest reflection on the sort of thing we ran into at the end of Chapter 4 (I think) and on the simplest paradigm cases of things which we can know about the past but not about the future (say, the fact that a certain egg, at a certain time, hits the floor and splatters in exactly the shape of Argentina) will get us to the crux of the matter.

To begin with, let's be a little more explicit about what the terms of the discussion are. The game here (once again) is to say how we manage to make the sorts of inferences that we do *from the present to times other than the present*, and to say why we *don't* manage to make certain *others*. And so we are going to need to start out with some sort of a ground rule as to what can be taken as unproblematically *given* to us, some ground rule about the sorts of *raw data* from which these inferences *proceed*. Let's settle, just for the moment, just to get the ball rolling, on a very crude one: take the raw data from which these inferences ought to be seen as proceeding to consist of what we were referring to at the end of Chapter 4 as the world's present directly surveyable condition.[2]

Good. Now, what I want to claim is that the crucial distinction between what we can know about the past and what we can know about the future (or rather, a very serviceable *first approximation* to that crucial distinction, an approximation which is going to turn out to be quite close to the mark, of which more later) runs as follows:

Everything we can know about the future (for example, the sun will rise tomorrow, the ice in the glass on the table in front of me will soon be melted, Henry is absolutely going to love *Lost Highway*) can in principle be deduced from nothing over and above the dynamical equations of motion and the probability-distribution which is uniform, on the standard measure, over the world's present directly surveyable condition. But a great deal of what we

2. The present directly surveyable condition of the world (you will remember) is the world's present macrocondition plus whatever (perhaps macroscopic) features of the present condition of the *brain* of the observer in question may be accessible to her by means of direct introspection. And granting that we know all that is, of course, granting us *much* too much: we don't have the *remotest idea* what the entirety of the present macrocondition of the world is, and we surely never will. Nonetheless, this assumption will serve well enough for the purposes of getting the structure of the arguments straight. Once that's done, we'll be able to refine it—in a perfectly straightforward way—to suit any doctrine we like about what it is we actually and currently and directly and unproblematically *know*.

take ourselves to know of the past (and it will be helpful for what follows to keep two very different sorts of examples simultaneously in mind here: (1) that certain particular eggs have on certain particular previous occasions splattered in exactly the shape of Argentina; and (2) that the entropy of the universe as a whole has previously been much lower than it currently is) *patently does not.*

The obviousness of all this notwithstanding, it has had an extraordinary way of escaping the attention even of investigators who were (in a manner of speaking) looking it straight in the face—investigators who were vividly aware (for example) of the fact that the overwhelming majority of trajectories passing, at any particular instant, through any non-maximal-entropy macrostate will have higher entropies on *both* temporal sides of that instant.[3] And *what that fact entails* (once again) is not only that almost the *entirety* of what we take ourselves to know of the past (that the entropy was lower, that certain eggs splattered in certain particular shapes, that the Roman Empire existed, and so on) *fails* to follow from the world's present macrocondition + the uniform microdistribution over that macrocondition + the laws of motion, but (rather) that it follows from all this that almost all of what we take ourselves to know about the past is almost certainly *false!* What will follow (more particularly) from the world's present macrocondition + the uniform microdistribution over that macrocondition + the laws of motion is (as we learned at the end of Chapter 4) that any book describing the Roman Empire is *far* more likely to have fluctuated out of molecular chaos than to have arisen as some sort of distant causal consequence of the *existence* of that empire; and no amount of *redundancy* among various such books, or among such books and archeological artifacts and whatever else you may be able to come up with, will change that one iota. Period.

▲▲▲ All right. Let's start to hook all this up with the terminology of the *methodological* asymmetries I mentioned before. Let's call inference procedures to other times which operate by plugging any available macroinformation about the present + the standard microstatistical rule into the

3. Even Reichenbach (in the passage I was discussing in footnote 17 of Chapter 4) gets this all bollixed up.

equations of motion *predictions* or *retrodictions*; and let's characterize all inference procedures which (for whatever reason) do *not* fit that description as relying on *records*. The claim, then, is that whatever we take ourselves to know of the future, or (more generally) whatever we take to be *knowable* of the future, is in principle ascertainable by means of *prediction*.[4] Some of what we take ourselves to know about the *past* (the past positions of the planets, for example) is no doubt similarly ascertainable by means of *retrodiction*—but far from all of it; rather little of it, in fact. Most of it we know by means of *records*.

And the puzzle is this: given that our direct empirical database presumably cannot *exceed* the world's present macrocondition (or the world's present *directly surveyable* condition; and no doubt, as I mentioned above, it is actually a great deal *smaller* than either of those), how can there *possibly* be reliable methods of inference to other times *other* than prediction and retrodiction? How *can* we know more of the past (as we in fact *do*) than can be deduced by means of retrodiction?

And at this point—unless all this is somehow put a stop to—a full-blown skeptical catastrophe is around the corner: retrodiction, after all, is going to count it as extraordinarily unlikely that the very experiments which we take to have *confirmed* classical mechanics in the *first* place ever actually *occurred*! And so one of the beliefs that the combination of classical mechanics and the uniform-over-the-present-macrocondition-of-the-world-probability-distribution would seem to *undermine* is the belief in classical mechanics *itself*! And so the reliability of prediction and retrodiction themselves (insofar as they both depend on the *truth* of the microscopic laws of motion) is patently going to require that there be some reliable technique of inference *other* than prediction and retrodiction. And the question now is what that other technique could possibly *involve*; the question is where the *extra information* that the technique will patently require can possibly be *coming* from.

4. Note, however, that it is certainly *not* being claimed here that we *routinely* arrive at our judgments about the future by means of anything along the lines of a *conscious application* of the method I have just described; only that whatever we take ourselves to know of the future could in principle *also* have been arrived at by means of an application of that method, and (moreover) that if it could be demonstrated of a certain thing that we take ourselves to know of the future that it could *not* have been arrived at by that method, we might begin to doubt that we know it after all.

▲▲▲ Let's start over.

Measuring devices (and this is the sort of thing that has long been recognized in discussions of measurement in the context of *quantum mechanics*) are *not* the sorts of systems whose states become reliably correlated with the states of the systems they are designed to measure merely in the event that they *interact* with those systems in the appropriate way. Indeed, insofar as the basic laws of nature are exclusively *dynamical* ones (of which more in a minute), there simply can't *be* any systems like that.

Here's what measuring devices *are:* measuring devices are the sorts of systems which reliably undergo some particular *transition*, when they interact in the appropriate way with the system they are designed to *measure*, only in the event that the measured system is (at the time of the interaction) in one or another of some particular collection of physical situations. The "record" which emerges from a measuring process is a *relation* between the conditions of the measuring device at the *two opposite temporal ends* of the interaction; the "record-bearing" conditions of measuring devices which obtain at one temporal end of such an interaction are reliable indicators of the situation of the measured system—*at* the time of the interaction—*only* in the event that the measuring device is in its *ready* condition (the condition, that is, in which the device is calibrated and plugged in and facing in the right direction and in every other respect all set to do its job) at the interaction's *other* temporal end. The sort of inference one makes from a *recording* is not from one time to a second in its future or past (as in prediction/retrodiction), but rather from *two* times to a *third* which lies *in between them*.

And note that inferences of this latter sort can be immensely more *powerful*, that they can be immensely more *informative*, than inferences of the predictive/retrodictive variety. Think (for example) of an isolated collection of billiard balls moving around on a frictionless table. And suppose that billiard ball number 5 (say) is currently at rest; and consider the question of whether or not, over the past ten seconds, billiard ball number 5 happens to have *collided* with any of the *other* billiard balls. The business of answering that question by means of retrodiction will of course require as input a *great deal* more information about the present—it will require (in particular) a complete catalogue of the present positions and velocities of all the *other* billiard balls in the collection. But note that the question can also be settled,

definitively, in the affirmative, merely by means of a single binary bit of information about the *past;* a bit of information to the effect that billiard ball number 5 was *moving* ten seconds ago.[5]

And the puzzle is about how it is that we ever manage to *come by* such information. It can't be by means of retrodiction/prediction (since, if that were the case, whatever other information that information could subsequently be *parlayed into* would necessarily *also* be of the predictive/retrodictive sort). It must be because we have a *record* of that other condition! But how is it that the ready condition of this *second* device (that is, the one whose *present* condition is the *record* of that *first* device's ready condition) is established? And so on (obviously) ad infinitum. There must (in order to get all this off the ground) be something we can be in a position to *assume* about some other time—something of which we *have* no record; something which cannot be inferred from the present by means of prediction/retrodiction—the mother (as it were) of all ready conditions. And this mother must be *prior in time* to everything of which we can potentially ever *have* a record, which is to say that it can be nothing other than the initial macrocondition of the universe as a whole.[6]

And so it turns out that *precisely* the thing that makes it the case that the second law of thermodynamics is (statistically) true throughout the entire history of the world is *also* the thing that makes it the case that we can have epistemic access to the past which is not of a predictive/retrodictive sort; the reason there can be records of the past and not of the future is nothing other than that it seems to us that our experience is confirmatory of a past-hypothesis but not of any future one.

And note, by the way, that all this leaves quite open the possibility that

5. Note also that it is nothing more than a verbal matter, in this billiard-ball example at least, which condition of the ball gets called the "ready" condition and which condition gets called the "record-bearing" one. Having the right sort of information about the billiard ball, or (more generally) about the measuring device, at some time other than the present, will make it possible to read its *present* state as a *record* of an occurrence at some time *between* those two.

6. There might be a temptation to think that the mother of all ready conditions—the ready condition (that is) about which we are prepared to make an *assumption,* the ready condition from which all the others are thereafter *inferred* by means of *records*—must be whatever condition of *our own brains* it is that ensures the reliability of our *sense perceptions.* But a little reflection will show that this can't possibly be right. The evidence of our senses can (after all) be *overridden,* on occasion, by other sorts of records—and there are (after all) records of events which occurred well before we were born!

there might turn out to *be* some as yet undiscovered future hypothesis (not, presumably, a hypothesis to the effect that the far-future state is a *low entropy* one, but that it is characterized by some simple *macroscopic* organization), of which our experience may yet prove confirmatory, and whereby we might yet learn to record or even to remember the *future* as well. That would make for a strange world, but perhaps not an altogether unintelligible one.[7]

Anyway, leaving that possibility aside, what seems to be the case is that everything we can know of the past and present and future history of the world can be deduced, in its entirety (as stated at the end of Chapter 4) from the following four elements: what we know of the world's present macrocondition—and of our own brains, perhaps; the standard microstatistical rule; the dynamical equations of motion; the past-hypothesis.

And the sorts of things that justify our *belief* in the past-hypothesis are (as before) precisely the sorts of things that typically justify our beliefs in *laws*. We believe in it (that is) not because it is *entailed*—or *can* be entailed—by any finite collection of particular empirical observations, but because of its conspicuous success (as in the case of the Napoleonic boots, or of the spatulas, or of the ice, or of countless billions of other things) in making *predictions* about how *future* particular observations are likely to *come out*,[8] and (more profoundly, perhaps) because it manages to render various of our *other* most fundamental convictions (about the veracity of our memories, and about the truth of the second law of thermodynamics, and about the accuracy of the dynamical equations of motion, and about the reliability of the techniques of prediction and retrodiction, and so on) *compatible* with one another.[9]

▲▲▲ One more detail—in connection with the attempt at the outset of this chapter to say precisely what it is that distinguishes the sort of epistemic

7. That it *isn't* unintelligible, or that (at any rate) it *perhaps* isn't unintelligible, is by no means a trivial matter. The sort of worry that jumps right out at you is that you might take steps to ensure that what you remember of the future does *not*, in the end, *occur*. And there is a famous old paper about just that, which is one of the prettiest and most imaginative little occasional pieces of physical reasoning I have ever come across, and which I heartily recommend to the reader: J. A. Wheeler and R. P. Feynman, "Interaction with the Absorber as the Mechanism of Radiation," *Reviews of Modern Physics* 17 (1945): 157–181.

8. This is the sort of thing I was talking about back on page 94, at the end of Chapter 4.

9. That is, it turns out to be precisely the thing we need in order to avert the full-blown skeptical catastrophe that was looming back on page 116.

access we have to the past from the sort of epistemic access we have to the future—now needs setting straight. That original attempt (it now appears) cannot possibly have gotten things exactly right, because some of what we can learn of the *past*, by means of records, can plainly also have implications about the *future*—implications which (more particularly) are altogether *over and above* what we can learn by means of *prediction*.

Think (for example) of the pseudo-Maxwellian demon we were talking about in the previous chapter—the one that rearranges the microconditions of isolated boxes of gas in such a way that at a certain particular time after those rearrangements have been completed, the entropy of the gas will begin to *decrease* spontaneously. And focus on a time *before* the entropy of the gas begins to decrease, but *after* the microscopic rearrangements have been completed—a time at which the gas and the demon are (therefore) no longer *interacting* with each other, a time (that is) at which the gas is a *fully isolated system*. And consider the probability-distribution that's uniform (on the standard measure) over that region of the phase space of the possible microconditions of the gas which is compatible with its *macrocondition* at that time—the probability-distribution (that is) which (in accord with our first attempt at saying precisely what it is that distinguishes the sort of epistemic access we have to the past from the sort of epistemic access we have to the future) is supposed to tell us everything we know, or *can* know, about how that gas is subsequently to be expected to *behave*. Well, the thing is, it *doesn't* tell us all that. *That* distribution (after all) is going to count it as overwhelmingly likely that the entropy of the gas—so long as it remains isolated—is *not* subsequently going to decease; and *we* (of course) *know better*.[10] What we know

10. And there's something worth pausing over and taking note of here (by the way) in connection with the relationship between these *pseudo*-Maxwellian demons and the *bona fide* ones.

Notwithstanding that both pseudo-Maxwellian demons and bona fide Maxwellian demons unquestionably amount to violations of the letter of the second law of thermodynamics, there is a certain conception of what it is that that law is genuinely *about*; there is a certain conception of what it is that that law is genuinely *getting at*—a conception which seems to have been on the minds of both Maxwell and his various critics, and which was discussed in some detail back in Chapter 5, and which has to do with getting energy out of the world by means of gross mechanical manipulations, according to which the pseudo-demons are not even remotely as fundamental or as catastrophic or as unspeakable a violation of that law as the *bona fide* ones are. And one of the things that the present considerations are turning up is that (as a matter of fact) there are *other* interesting conceptions of the spirit of that law on which precisely the *reverse* is true. The operations of the *bona fide* demons (after all) are perfectly in accord with the principle that the future macroscopic behaviors of isolated macroscopic systems are more or less accurately given

(more particularly) is that after a certain definite time has elapsed, the gas is spontaneously going to begin to (say) *contract*. And the *way* we know that, the thing we *infer* it from, is a *memory*, a *recording*, of the fact that this demon and this gas have previously (and in the appropriate way) *interacted* with each other.

Or imagine that at a certain time t we come upon a box with a gas in it. And suppose (moreover) that the box has a *sign* on it. And suppose that the sign reads: "Just before t, with a microscope, I observed that gas particle number 2874 was located at point ($x = 39.7$, $y = 12.1$, $z = 5$).—Sidney." And suppose that Sidney is someone I know to be an invariably serious and trustworthy person, and an excellent microscopist. And the thing is that all this manifestly puts me in a position to infer the outcome of a *new* measurement of the position of particle 2874—a measurement which is to be carried out just *after* t—with enormously greater accuracy than I could have merely on the basis of the probability-distribution which is uniform (on the standard measure) over that region of the phase space of the box + the gas + the sign which is compatible with its present macrocondition. And the origin of this extra information is (once again) manifestly in the possibility of reading the sign as the *record* of an event in the *past*.[11]

---

by the probability-distribution that's uniform (on the standard measure) over the microconditions compatible with their present macroconditions, but what's just emerged here is that the operations of the *pseudo*-demons are radically *in violation* of that principle!

And as a matter of fact (if you stop and think about it) the considerations of Chapter 5 entail that there is nothing whatsoever in the classical equations of motion which stands in the way of there being (even) *super*-Maxwellian demons, which violate the second law on *both* of the above conceptions of what that law is trying to say.

What all this is positively crying out for (of course) is nothing along the lines of a decision as to what the genuine essence of the second law finally turns out to *be*, but (rather) an *improvement* of that law of the sort that I was pining after in the last two paragraphs of Chapter 5—a *better* and more *explicit* and absolutely *universal* version of that law in which letter and essence are finally *one*.

11. And (come to think of it) if the complete dynamical theory of the world is deterministic, then anything *whatsoever* that constrains the condition of the world in the past necessarily *also*, *somehow*, constrains the condition of the world in the *future*. And if (moreover) the complete dynamical theory of the world happens to be in accord with the premises of *Liouville's* theorem, then there is a perfectly straightforward sense in which anything that constrains the condition of the world in the past necessarily also constrains the condition of the world in the future *to exactly the same degree*. And so the sort of "knowing" we have in mind when we speak of "knowing more of the past than we do of the future"—if that sort of talk is to make any *sense*—must amount to something *other*, something more *specific*, than our merely being in a position to rattle off some constraints on the condition of the world. The idea (presumably) is that when we

And so our initial stab at saying what it is that distinguishes the sort of epistemic access we have to the past from the sort of epistemic access we have to the future can certainly not (once again) have gotten things exactly right.

The right thing (insofar as I know) is this:

Start with a probability-distribution which is uniform—on the standard measure—over the world's present macrocondition. Conditionalize that distribution on all we take ourselves to know of the world's entire macroscopic past history (and this will amount to precisely the same thing—if you think it over—as conditionalizing it on the *past-hypothesis*). Then evolve this conditionalized present-distribution, by means of the equations of motion, into the future.

This will yield (among other information) everything we take ourselves to know of the future.

Conversely:

Start with the same uniform probability-distribution over the present macrocondition. Conditionalize this distribution on everything we take ourselves to know of the world's entire macroscopic *future* history (and this will amount to very nearly—but not quite—no conditionalization at all). Then evolve *this* conditionalized present-distribution, by means of the equations of motion, into the *past*.

This will yield immensely *less* than we take ourselves to know of the past.

---

speak of ourselves as "knowing something about the past," we are taking ourselves to be in a position to rattle off the sorts of constraints which pertain to some previous condition of a relatively small and unisolated *subsystem* of the world (Napoleon, say, or Woody Guthrie, or Greece), and which can be expressed in a relatively simple, natural, straightforward, everyday sort of language. And there is patently nothing whatsoever either in the determinism of the dynamical laws of nature or in the Liouville theorem which requires that constraints *like that* about the past necessarily also amount to constraints like that about the *future*.

Think (for example) of the billiard balls we were talking about before. And call the proposition that billiard ball number 5 is involved in a collision within a certain particular time-interval—the past ten seconds, say—proposition *P*. A proposition like *P*, then—together with the dynamical equations of motion—is inescapably going to amount to a *constraint* on the possible physical conditions of that collection of balls at *every one* of the temporal instants *outside* of the interval in question *as well*; and (moreover) the *information content* of *any individual one* of those constraints (that is, the information content of the constraints on any individual one of those other *times*) is inescapably going to be equal to the information content of proposition *P itself*.

The rub (if you think about it) is just that those other constraints are generally going to take the form of more or less unimaginably complicated correlations among the values of certain of the physical properties of nothing less than *every last one* of the billiard balls in the *collection*, which is to say that they are generally *not* going to amount to the sorts of propositions in which human beings are even structurally capable of taking an interest.

▲▲▲ That (at any rate) is what seems to me to be the right and final way of putting things in the classical context. But once again, there is a speculative history concerning these matters which (notwithstanding that it is in many ways a misguided history) deserves some of our attention. The speculations in question here were undertaken (remember) without the benefit of any clearly enunciated distinction between what we can know of the past and future, or between the methods of prediction/retrodiction and recording; and consequently, they were never able to focus squarely (as we have here) on the question of what more *fundamental* asymmetries might *account* for those distinctions. What they were all trying to do was just to construct an argument to the effect that records (whatever records *are*, precisely) must necessarily and as a matter of fundamental principle be records of the entropic *past*.[12]

Reichenbach (for example) suggests that we imagine encountering a system which is more or less isolated—at present—and which is currently in a state of non-maximal entropy relative to its current gross constraints (a half-melted chunk of ice floating in a glass of warm water, for example). Our speculations about the past and future of this system must now (says Reichenbach) run as follows: if the system in question is permanently isolated, its present state must be counted as an enormously improbable *fluctuation*; but the assumption that some *interaction with the external world* in the near past or future (somebody's dropping an *entirely* unmelted ice cube into the glass of water, for example)—an interaction which leaves the system in a macrocondition whose entropy is even *lower* than the entropy of the system's condition at *present*—will render that present condition overwhelmingly *probable*. Moreover, our general background knowledge of the world will assign a not-particularly-low a priori probability of such an interaction's taking place. And in light of that, the present non-maximal entropy condition of this system will rightly be viewed, under normal circumstances, as a *record* of the actual *occurrence* of such an interaction. And (this is the punch line) that this interaction must necessarily lie in the *past* of its record will

12. And note that this is a conclusion which, if the above considerations are correct, is not even *true*. What emerges from *those* considerations is that the fact that there happen not (in *our* world, insofar as we are currently *aware*) to be records of the entropic *future* is an utterly contingent matter: it has to do with the fact that there happens to be a past-hypothesis and there happens *not* (insofar as we are currently aware) to be a *future* one.

now follow from the general statistical truth of the second law; it will follow (that is) from the fact that a uniform microdistribution over the initial very-low-entropy macrocondition of the world amounts to a satisfactory theory of the history of the universe. And all this seems to me to be precisely on the mark—*given the assumption* that record-bearing conditions (of *every imaginable variety*) are invariably *non-maximal-entropy* ones. But *that* just seems wildly *wrong*. There just isn't *anything whatsoever* (if you stop and think it over) that's somehow part and parcel of something's being a *record-bearing* condition to the effect that it should (at present) be *irreversibly changing*—or changing *at all*. Suppose (for example) that I come back to my apartment in the evening and find it in precisely the same stable and disheveled condition I left it in that morning. Is there any reason in the world not to read that condition—in the *evening*—as a *record* of the fact that nobody broke into my apartment during the day and *cleaned it up*?

The only other interesting stab at this I know of comes from Paul Horwich in his book *Asymmetries in Time*. It goes like this.

The idealized account of measuring devices which we've been making do with thus far supposes that the transition from the ready state to the record-bearing state occurs *only* when the sort of event that the device in question is designed to *look for* occurs. We presumed, for example, that a stationary billiard ball can only come to be moving later on in the event that something *collides* with it in the interim.

But Horwich correctly points out that this sort of thing is not, in general, going to be exactly true. It isn't *impossible*, for example, for an initially stationary billiard ball to come to be moving by means of a *fluctuation*. What it surely *is*, however, is extraordinarily *unlikely*; and it is in virtue of *that* (the fact that devices can be constructed for which "spurious" readings, though possible, will be rare) that there can be such things in the world as reliable recordings.

Note, however, that if we follow the evolution of the world *backward* in time, bizarre and unpredictable fluctuations (a stationary billiard ball suddenly cools off and starts to move, for example) will be no less ubiquitous than *approaches to equilibrium* are as we follow the world's evolution *forward* in time. And *this* (that is, that the bizarre and the unpredictable will be the rule as we follow the evolution of the world toward its *low*-entropy extremum)

will (so Horwich thinks) render reliable recording devices for the *future* (that is, devices whose record-bearing states *precede* their ready states) impossible. The idea is that the construction of a recording device of this latter type for which spurious readings could somehow be guaranteed to be rare is (given what we have just observed about how the world looks if we follow its evolution *backward* in entropic time) manifestly out of the question.

But a little further thought will show all this to be silliness. The point is not to get carried away. It is emphatically *not* the case that, going backward in time, *anything* can happen. Strange and (within limits) unpredictable things can happen, but *only* those things which, seen in the other time-direction, are in accord with the second law!

If, for example, a billiard ball is known to be stationary at one entropic time and moving at a later one, then (*whichever* of these two states gets called the "ready" state, and whichever gets called the "record-bearing" one) a collision can be inferred. Period.

2. INTERVENTION

There's one further fundamental conviction we have about the difference between the past and the future, which is that the future *depends on what happens now*—that the future depends on what we *do* now—in a way that the past does not. And all the apparatus we need for getting to the bottom of that is (I think) now in place.

The first thing that needs saying (I suppose) is that there is nothing whatsoever in the way of a tension between (on the one hand) the conviction that the future depends on what we do now and (on the other) the proposition that the fundamental and universal equations of motion of the entire physical world (including us, of course) are fully *deterministic*. What we have in mind when we say that the future depends on what we do now is certainly not that the *actual* future can somehow be anything other than what it is actually going to *be*, or that there's anything *about* that future that isn't completely and irrevocably *nailed down* by the world's *initial condition*, but merely that *if* (contrary to fact) we *were* to be doing other than we actually happen to be doing, at present, then the *future* would be something other than it actually *is*. What we think (that is) is that the future counterfactually *depends* on the present—and (moreover) we think that the future depends

on the present in a way that the past emphatically does *not*. We think (for example) that the decisions that the President of the United States makes today somehow *matter* vis-à-vis the question of whether there will be a nuclear war tomorrow in a way that they *don't* vis-à-vis the question of whether there *was* a nuclear war *yesterday*. And there is manifestly (again) a *question*—in light of the invariance under time-reversal of the fundamental dynamical equations of motion—about how that can possibly *be*.

And the answer to that question is intimately bound up (I think) with the fact that we live in the sort of world whose simplest and most informative description—the description at the very end of Chapter 4—involves a *past*-hypothesis but no *future* one. The idea (more particularly) is that that fact, that *asymmetry*, can be parlayed into an argument to the effect that the present determinants of the past are (as it were) enormously less *amenable to our control* than the present determinants of the *future* are.

Think (to begin with) of the collection of billiard balls we were talking about before. And suppose that some particular one of those balls (ball number 5, say) is currently stationary. And suppose (and this is what's going to stand in—in the context of this extremely simple example—for a *past-hypothesis*) that that same ball is somehow known to have been *moving* ten seconds ago.

What we learned about that sort of a collection of balls in the previous section (you will remember) was that whereas (on the one hand) whether or not ball number 5 will be involved in a collision over the next ten seconds is determined by the present condition of the *entire collection* of balls, whether or not ball number 5 *has been* involved in a collision over the *past* ten seconds is (on the other) unambiguously determined—under these circumstances[13]—by ball number 5's present condition *alone*.

And this is something it will be worth taking the trouble to put in one or two slightly different ways.

One of the things this means is that whereas (on the one hand) there are patently *any number* of hypothetical alterations of the present condition of the balls in this set—*whatever* that condition might happen to be—which would alter the facts about whether or not ball number 5 is to be involved in

13. That is, given the information we have about the condition of this ball *ten seconds ago*.

a collision over the next ten seconds,[14] there can (on the other) be *no* hypothetical alterations in the present condition of this set of balls, unless they involve hypothetical alterations in the present velocity of ball number 5 *itself*, which would alter the facts about whether or not ball number 5 *had been* involved in a collision over the *past* ten seconds.[15]

And *another* of the things it means (and this is the one that's going to be the most useful for our purposes here) is that whereas (on the one hand) there are perfectly imaginable present conditions of this collection of balls in which certain small hypothetical alterations of (say) the present velocity of ball number 12 would alter the facts about whether or not ball number 5 is to be involved in a collision over the next ten seconds,[16] and there are perfectly imaginable present conditions of this collection of balls in which certain small hypothetical alterations of the present position of ball number 2 would alter the facts about whether or not ball number 5 is to be involved in a collision over the next ten seconds,[17] and there are (in short) perfectly imaginable present conditions of this collection of balls in which certain small hypothetical alterations of *any physical feature you choose* of *any particular one of these balls you like* would alter the facts about whether or not ball number 5 is to be involved in a collision over the next ten seconds, there are (on the other hand) *no* imaginable present conditions of this collection of balls at all (so long as those conditions are compatible with the proposition that ball number 5 is currently stationary, and so long as it is taken for granted that ball number 5 was moving ten seconds ago) in which any hypothetical alterations whatsoever in the present conditions of any of these balls

14. That is, *whatever* the present condition of these balls is, it will manifestly be possible to specify hypothetical alterations of (say) nothing over and above present conditions of ball number 7, or of nothing over and above the present conditions of balls 25 and 32, or of nothing over and above the present conditions of balls 12 through 29, or of nothing over and above the present conditions of any subset of this collection of balls *whatsoever*—which will alter the facts about whether or not particle number 5 is involved in a collision over the next ten seconds.

15. There can be no such alterations (that is) so long as that actual present condition in question is (as stipulated above) one in which ball number 5 is stationary, and so long as we hold it fixed throughout all the *hypothetical alterations* of that present condition that ball number 5 was moving ten seconds ago.

16. Conditions (for example) in which ball number 12 is very nearly (but not quite) on a collision course with ball number 5—or conditions in which ball number 12 is very nearly (but not quite) on a collision course with some *other* ball which is *itself* very nearly (but not quite) on a collision course with ball number 5.

17. Conditions (say) in which ball number 2 is very nearly (but not quite) *in the path* of ball number 5.

other than ball number 5 itself would alter the facts about whether or not ball number 5 *was* involved in a collision over the *past* ten seconds.

And so there are (as it were) a far wider variety of potentially available *routes to influence* over the future of the ball in question here, there are a far wider variety of what we might call *causal handles* on the future of the ball in question here, under these circumstances, than there are on its past.

All right. Let's jack all this up to cases of worlds more or less like *our own*—worlds (that is) in which there are people and buildings and cities and oceans and planets and galaxies and what have you, all of which behave more or less as we're *used* to such things behaving. The thing we want to *find out* about (remember) is precisely *which ones* of the above-mentioned sorts of handles we are capable—under the right circumstances—of *getting a hold of*. And any systematic investigation of that sort of a question is going to have to start out (insofar as I can tell) with some primitive and un-argued-for and not-to-be-further-analyzed conception of which particular features of the present condition of the world it is that are to be thought of as falling under our (as it were) *direct* and *unproblematical* and *unmediated* control. Those features might be (say) the positions of my hands and feet and fingers and toes, or the tensions in various of my muscles, or the electrical excitations (or lack of them) in various of my motor neurons, or even the conditions of various regions of my *brain*. And a little reflection will show that it isn't going to *matter* very much, vis-à-vis the question of how the analysis is ultimately going to *come out*, *which particular one* of the above candidates for the directly controllable we happen to *pick*. Whatever gets decided about that, the crux of the matter is going to be that (in the first place) the part of the present condition of the world over which we can reasonably think of ourselves as in the above sort of direct and unproblematical control is *exceedingly tiny*, and that (in the second place) for more or less *any present feature of the world you can think of* and for more or less any *future* feature of the world you can think of there are going to be a host of perfectly imaginable worlds more or less like our own in which the *present* feature in question amounts to a *causal handle* on the *future* one, and that (in the third place) there are exceedingly *few* present features of the world which will amount—on *any* imaginable world more or less like our own—to a causal handle on any particular feature of the *past*.

Let's spell all this out in a tiny bit more detail. The fundamental point is

that in worlds like ours, the right procedures for making inferences to other times from the present (let's call them the normal procedures of inference, or the NPI for short) are the ones spelled out at the very end of Chapter 4. Indeed, that the right procedures for making inferences to other times from the present—in a certain hypothetical world—are the ones at the end of Chapter 4 is part and parcel of what we *mean* when we speak of that world as being "like ours." And there are patently worlds like that in which the NPI translate relatively small hypothetical differences in (say) the present position of the right index finger of the President of the United States into the difference between a certain thermonuclear device's exploding or not exploding two minutes down the road.[18] And there are other worlds like that in which the NPI translate relatively small hypothetical differences in the present position of the *left foot* of the President of the United States into the difference between a certain thermonuclear device's exploding or not exploding two minutes down the road.[19] And there are still *other* worlds like that in which the NPI translate small hypothetical differences in the present position of the left foot of the President of the United States into hypothetical differences in the channel that a certain television set is going to be tuned to five seconds down the road.[20] And there are (in short—and as I mentioned above) worlds more or less like this one in which the NPI translate almost any hypothetical *present* difference you can think of into almost any hypothetical *future* difference you can think of.

But note that the situation is utterly and absolutely different with respect to the past. There are (for example) no worlds *at all*, even *remotely* like our own,

18. Maybe this is worth saying very carefully. What we mean (to begin with) by "a world like ours" is (again) a world in which there are people and buildings and oceans and planets and (importantly) whose simplest and most compact and most informative description is the one at the end of Chapter 4. And what's being *claimed* here is (in the first place) that there is some perfectly imaginable present condition of a world like that (a condition in which—among other things—the right hand of the President of the United States is somewhere in the immediate vicinity of the button) for which the NPI entail that there is going to be a thermonuclear explosion two minutes down the road, and (in the *second* place) that there is some *other* perfectly imaginable present condition of a world like that, one that differs from the first in only a small way, and almost exclusively in terms of the position of the President's right index finger, for which the NPI entail that there is *not* going to be a thermonuclear explosion two minutes down the road.

19. *These*, of course, will be worlds in which the *left foot* of the President of the United States is somewhere in the immediate vicinity of the button.

20. You get the idea.

in which the NPI translate small hypothetical present differences in the present position of anybody's finger into the difference between a certain thermonuclear device's exploding or not exploding two minutes *ago*. And that (as I said before) is precisely because there is a past-hypothesis and not a future one. That (to put it another way) is because there are—vis-à-vis such things as the *past* explosions of thermonuclear devices (or the lack of them)—such things as *records*, as *memories*. And it is (as we've seen) part and parcel of the *logic* of those sorts of things that there just can't *be* any hypothetical present differences in a world like ours that the NPI are going to translate into the difference between a certain thermonuclear device's exploding or not exploding two minutes ago—unless (of course) those present differences *specifically* involve whatever *records* or *memories* there may be in the world, at present, of whether or not any such explosion actually *took place* two minutes ago.

And those sorts of records are manifestly *not* things that can by any stretch of the imagination be thought of as among the features of the present condition of the world that fall under our *direct* and *unmediated* and *unproblematical control*.

And I am presuming here that however the details of the right method of evaluating the truth-values of counterfactuals come out (and how those details *do* come out is still very much an open question), that method is going to have to *respect* and to *abide* by and to be in *accord* with all the consequences of the NPI.

And it follows—if all this is right—that the future does indeed counterfactually depend on what we do now, and that the past (for all our intents and purposes) does not.[21]

21. Of course, none of this means that the past is as a general matter counterfactually *independent* of the present. It patently *isn't*. It ought to have been clear from the outset (if we had bothered to think it through) that if the present were different from what it actually is in a sufficiently radical way—if (say) the present configuration of the entire material contents of the universe were a gigantic scale model of Bozo the Clown—then the past would surely have been radically different too. But that sort of thing can be of no practical interest whatsoever to creatures like us—for whom (at best) only the minutest details of the present condition of the world, or rather, only a minute *collection* of the minutest details of the present condition of the world, can reasonably be taken to be under our control. The point that's crucial for creatures like *us* (and precisely *this* is the point of our whole discussion of intervention) is that whereas almost *any* particular feature of the present can turn out—under the right circumstances—to be counterfactually decisive vis-à-vis almost any particular feature of the *future*, *precious few* particular features of the present condition of the world can turn out, in worlds anything at all like the one *we* live in, to be counterfactually decisive vis-à-vis any particular feature of the *past*.

QUANTUM MECHANICS

1. THE BACKGROUND

Newtonian mechanics—as I mentioned at the outset—happens not to be the mechanics of our world; the mechanics of our world (insofar as anybody can tell at present) is quantum mechanics.

The empirical predictions of those two theories more or less coincide, of course, insofar as the sorts of things that Newtonian mechanics manifestly gets right (things like the motions of planets, or of baseballs, or even—under certain circumstances—of molecules) are concerned, but the fundamental pictures they present of the space of possible physical states and of the evolutions of those states over time are altogether different from each other.

And so a question very naturally comes up as to whether and how the sort of universal Newtonian statistical mechanics we have been working out over the past six chapters can (as it were) be *adapted* to *quantum theory*.

And the conventional wisdom is that this sort of adaptation is (to begin with) a perfectly accomplishable thing, and that (as a matter of fact, and particularly in light of the relative closeness of the quantum and the Newtonian theories to each other insofar as things like the motions of molecules are concerned) the whole business turns out to be an eminently simple and straightforward matter of *translation*.

It (or rather, what seems to me to be the best available *rational reconstruction* of it in terms of the cleaner and more precise sort of vocabulary we have been working out over the course of this book) goes like this:

To begin with, the dynamical laws that govern the evolutions of the instantaneous states of quantum-mechanical systems in time (or rather, the dynamical laws that are *usually taken* to govern those evolutions—of which

more later) involve only *first derivatives*.[1] And one of the things that entails is that the instantaneous states of quantum-mechanical systems are invariably also the *complete dynamical conditions* of those systems.[2] And one of the things *that* entails (if you think about it for a minute) is that the dynamical laws that govern the evolutions of quantum states in time cannot possibly be invariant under *time-reversal*.[3] And yet (and this is the main point) the quantum-mechanical equations of motion *do* have the sort of *partial* invariance under time-reversal—the invariance under time-reversal insofar as the *positions of particles* are concerned—that we talked about in Chapter 1. And equations like that (as we've already seen) are going to present exactly as much of a problem vis-à-vis the time-directedness of our everyday macroscopic experience as the *Newtonian* equations do—and (moreover) they're going to present it in more or less exactly the same *way*.

And (as to the question of what might imaginably be *done* about that problem) you can run a quantum-mechanical argument—very much along the lines of Boltzmann's classical ones, very much (as a matter of fact) *parasitic* on Boltzmann's classical ones—to the effect that the familiar regularities in the evolutions of the macroconditions of thermodynamic systems toward the future can very plausibly be deduced from the initial macrocondition of the system in question together with a quantum-mechanical version of the statistical postulate (in which the standard measure over the classical phase space is replaced by an analogous one over sets of possible

1. This contrasts sharply, of course, with the equations that govern the evolutions of *Newtonian* systems—which have a d^2/dt^2 in them.

2. And this contrasts sharply with the Newtonian case too. *There* (remember) states consist simply of specifications of positions, whereas dynamical conditions consist of specifications of both positions and velocities.

3. The idea is this: suppose that the instantaneous microscopic state of a certain physical system at time t is also that system's complete dynamical condition at t, and suppose that the dynamical equations of motion of that system are invariant under time-reversal. Then whatever it is that those equations entail about times *other* than t is patently going to have no alternative whatsoever but to be *symmetrical* about t. Suppose (moreover) that the equations of the motion of this system are invariant under time-*translations*—which is the case (by the way) of every single one of the candidates for a fundamental dynamical equation of the evolution of the world that anybody has taken seriously since the scientific revolution of the Renaissance. *Then* (if you think it over) the state of this system is going to have no alternative but to be entirely *unchanging* in time. And so any theory for which instantaneous states are also invariably complete dynamical conditions, and for which the equations of motion are invariant under time-reversal, and for which the equations of motion are invariant under time-*translation*, is necessarily a theory according to which *nothing ever happens*.

quantum states) together with what are usually taken to be the correct quantum-mechanical equations of motion. And there will (of course) be more or less the same sorts of *reversibility objections* to all this as there were in the classical case—and those objections will elicit more or less the same sorts of *precisifications* of the Boltzmannian scenarios as they did in the classical case—and those precisifications will in turn run into more or less the same sorts of problems about the *past* as they did in the classical case. And the point to which all the standard sorts of attempts at *fixing up those problems* will inexorably *lead* is precisely the set of fundamental postulates of universal statistical mechanics discussed at the end of Chapter 4—with the classical equations of motion replaced (of course) with what are usually taken to be correct quantum-mechanical ones, and with the classical *statistical postulate* replaced by its above-mentioned quantum-mechanical *correlate*.

And that (so the story goes) is pretty much all there is to it.[4]

▲▲▲ And the thing I want to think through in this last chapter is whether or not there might be more to the story than that. The thing I want to think through in this last chapter is whether or not the transition to quantum theories might somehow (while we're at it) accommodate a much more radical

4. The *empirical thermodynamical consequences* of the classical and the quantum-mechanical versions of statistical mechanics are (of course) going to *differ*—but those differences (as with the differences between classical and quantum mechanics *themselves*) turn out not to amount to much except in fairly unfamiliar sorts of circumstances—at very low *temperatures*, say, or in very small *containers*, or on very oddly shaped *energy hypersurfaces*, or what have you. Elsewhere (which is to say—insofar as things like the spreading of ordinary sorts of smoke and the cooling of ordinary sorts of soup are concerned) classical mechanics and quantum mechanics (as I mentioned above) more or less reduce to each other, and (consequently) the standard formulations of classical and quantum-*statistical mechanics* do too, and (consequently) the empirical contents of classical and quantum *thermodynamics* do *too*.

And one other thing deserves mention here. There is a very deeply entrenched tradition in the physical literature of attributing whatever empirical differences there are between quantum and classical statistical mechanics at least in part to differences between quantum mechanics and classical mechanics on the question of Haecceisstism. But as a matter of fact (as discussed at some length in Chapter 3), classical statistical mechanics turns out (on closer examination) *not* to be committed to *any position whatsoever* on that question, and so quantum mechanics and classical mechanics cannot possibly be in any *disagreement* about it, and so whatever empirical differences there are between quantum and classical statistical mechanics are necessarily matters of *physical theory* (which is to say, they are necessarily matters of the structure of the space of possible physical states, and of the statistical-mechanical measure over that space, and of the dynamics of the evolutions of those states in time) and *not* of *metaphysical doctrine*, and how (come to think of it) can it possibly have been otherwise?

and more interesting sort of transformation of the structure of statistical mechanics than that.

The thing (remember) is that there's a difficult *problem* at the *foundations* of quantum theory. And nobody is quite certain—just yet—precisely what ought to be *done* about that problem. And it might well turn out that the dynamical laws of the evolutions of the states of quantum-mechanical systems in time are *not* (in fact) what they have usually been taken to be. And that might change everything.

But this is getting ahead of ourselves. Let's talk some (to start with) about precisely what the problem at the foundations of quantum mechanics *is*.[5]

2. THE MEASUREMENT PROBLEM

Here are some true stories about experiments with electrons.

The experiments all involve measurements of two perpendicular components of the intrinsic angular momenta, of two perpendicular components of what are usually referred to as the "spins" of electrons.

Let's call them the *x*-spin and the *y*-spin.

It happens to be an empirical fact (insofar as we currently know, at any rate) that the *x*-spins of electrons can assume only one of two possible values, and the same goes for *y*-spins.

Let's call those values +1 and −1.

The measurement of *x*-spins and *y*-spins is something which can be accomplished, with currently available technologies, with considerable ease and with considerable accuracy. The usual sorts of *x*-spin and *y*-spin measuring devices (which will henceforth be referred to here as "*x*-boxes" and "*y*-boxes") work by altering the direction of motion of the measured electron on the basis of the value of its measured spin component, so that the value of that spin component can be determined later on by a simple measurement of the electron's *position*. (See Figure 7.1.)

Another empirical fact about electrons is that there are as a rule no corre-

5. What follows here, by the way, is not going to amount to a particularly deep or particularly exhaustive account of that problem. And readers who are not already familiar with that problem, and with the various proposed strategies for solving it, are (I think) going to need to hear more. And perhaps it's worth mentioning (in this connection) that I wrote a book about all that some years ago, which is called *Quantum Mechanics and Experience*, and from which the present remarks have been adapted.

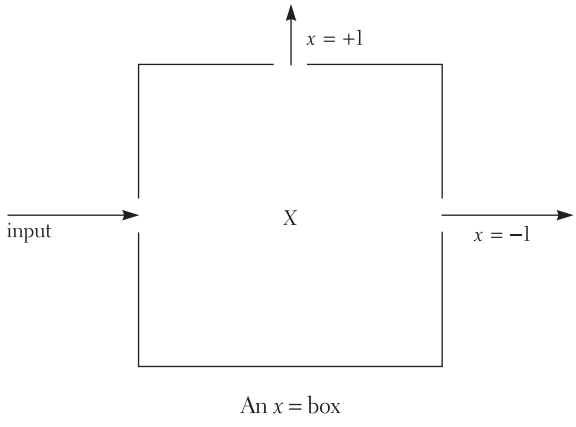


Figure 7.1

lations whatsoever between their *x*-spin values and their *y*-spin values: of any large collection of, for example, *x*-spin = +1 electrons, all of which are fed into the left aperture of a *y*-box, precisely half (statistically speaking) will emerge through the *y*-spin = +1 aperture, and half will emerge through the *y*-spin = −1 aperture; and the same goes for *x*-spin = −1 electrons fed into the left aperture of a *y*-box, and the same goes for *y*-spin = +1 and *y*-spin = −1 electrons fed into *x*-boxes.

And *another* empirical fact about electrons, and an extremely important one for our purposes here, one that is worth discussing in some detail, is that a measurement of the *x*-spin of an electron can disrupt the value of its *y*-spin, and that a measurement of the *y*-spin of an electron can disrupt the value of its *x*-spin, in what appears to be a completely uncontrollable way.

If, for example, a measurement of *y*-spin is carried out on any large collection of electrons *in between* two measurements of their *x*-spins (as in Figure 7.2), what invariably happens is that the *y*-spin measurement changes the *x*-spin values of half (statistically speaking, again) of the electrons that pass through it, and leaves the *x*-spin values of the other half unchanged.

No one has ever been able to design a measurement of *y*-spin which has anything other than precisely that effect on *x*-spin values, and no one has ever been able to identify any physical property or any combination of physical properties of the individual electrons in such collections which deter-

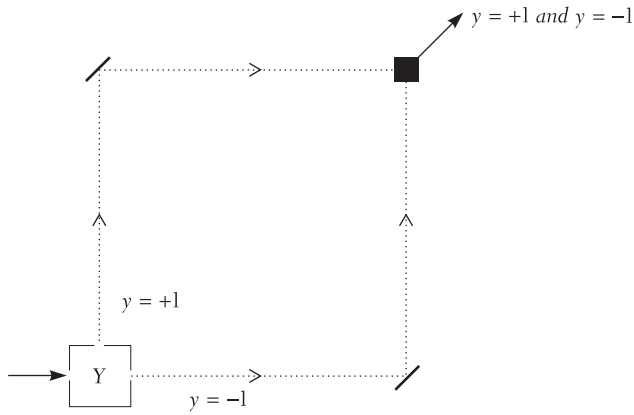


Figure 7.2

mines *which* of them get their *x*-spins changed in the course of having their *y*-spins measured and which *don't*.

What the official doctrine has to say about these matters (and these seem like very innocent and very reasonable inferences from the experimental data) is that there can in principle be no such thing as a *y*-spin measurement which has anything other than precisely that effect on *x*-spin values, and that which electrons get their *x*-spins changed by measurements of their *y*-spins and which don't is a matter of pure dynamical *chance*, that (in other words) the laws which govern those changes simply *fail* to be *deterministic*.

If (by the way) there can in principle be no such thing as a measurement of *x*-spin which fails to disrupt uncontrollably the value of *y*-spin, and if there can in principle be no such thing as a measurement of *y*-spin which fails to disrupt uncontrollably the value of *x*-spin, then, patently, there can as a matter of principle be no way of ascertaining *both* the value of the *x*-spin *and* the value of the *y*-spin of any particular electron at any particular moment. And that fact is an example (but only one among literally infinitely many) of the *uncertainty principle*: measurable physical properties like *x*-spin and *y*-spin are said to be "incompatible" with each other, since measurements of one will always (so far as we know) uncontrollably disrupt the other.

▲▲▲ Let's get in deeper.

Consider the rather complicated device shown in Figure 7.3. In one cor-

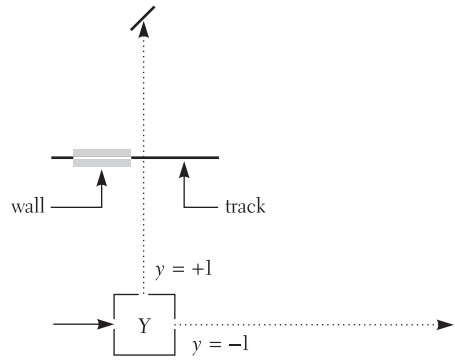


Figure 7.3

ner there's a *y*-box. *Y*-spin = +1 electrons emerge from that box along the route labeled "*y* = +1," and at a certain point on that route there's a "mirror" or a "reflecting wall" which changes the direction of motion of the electron but doesn't change anything else about it (more particularly, it doesn't change the value of the *y*-spin of an electron that bounces off it) as shown. And similarly for *y*-spin = −1 electrons.

At the point where the two routes reconverge, there's a "black box," which also changes the directions of the motions of electrons, without altering the values of their *y*-spins in the process, in such a way as to make the two routes coincide after they pass through it.

Let's do an experiment.

Suppose that we feed a large collection of *x*-spin = +1 electrons, one at a time, into the *y*-box; and then, as they emerge from the apparatus at "*y* = +1 *and y* = −1," we measure their *x*-spins. What sorts of results should we expect? Well, our previous experience informs us that half of such electrons (statistically speaking) will turn out to have *y*-spin = +1, and so will take route "*y* = +1" through the apparatus; and half of them will turn out to have *y*-spin = −1, and so will take route "*y* = −1" through the apparatus. Consider the first half. Since nothing that those electrons will run into in between the *y*-box and "*y* = +1 *and y* = −1" can have any effect on their *y*-spin values, they will all emerge from the apparatus as *y*-spin = +1 electrons, and consequently (our experience informs us here again) 50 percent of *them* will turn out to have *x*-spin = +1 and 50 percent will turn out to

have x-spin $= -1$. The second half, by contrast, will all emerge as y-spin $= -1$ electrons, but of course their *x-spin* statistics will be precisely the same. Putting all this together, it follows that of any large set of x-spin $= +1$ electrons which are fed into this apparatus, half should be found at the end to have x-spin $= +1$ and half should be found at the end to have x-spin $= -1$.

All that seems absolutely cut and dried.

But the funny thing is that when you actually go out and try this experiment, what happens is that exactly 100 percent of the x-spin $= +1$ electrons that initially get fed into this apparatus (one at a time, mind you) come out with x-spin $= +1$ at the end.

This is very odd. It's hard to imagine what can possibly be going on. Perhaps things will get a bit clearer if we do another experiment. Suppose that we rig up a small, movable, electron-stopping wall that can be slid, at will, in and out of, say, route "$y = +1$" (see Figure 7.3). When the wall is "out," we have precisely our earlier apparatus; but when the wall is "in," all electrons moving along "$y = +1$" get stopped, and only those moving along "$y = -1$" get through to "$y = +1$ *and* $y = -1$."

What should we expect to happen when we slide the wall in? Well, to begin with, the overall output of electrons at "$y = +1$ *and* $y = -1$" ought to go down by 50 percent, since the input x-spin $= +1$ electrons ought to be half y-spin $= +1$ and half y-spin $= -1$, and the former shouldn't now be getting through. What about the x-spin statistics of the remaining 50 percent? Well, when the wall is out, 100 percent of the x-spin $= +1$ electrons initially fed in end up as x-spin $= +1$ electrons. That means that all the electrons that take route "$y = +1$" end up with x-spin $= +1$ *and* that all the electrons that take route "$y = -1$" end up with x-spin $= +1$, and since we can easily verify that whether the wall is in or out of route "$y = +1$" can have no effect whatsoever on the x-spins of electrons traveling along route "$y = -1$," that implies that those remaining 50 percent should be all x-spin $= +1$.

What actually happens when we go and do the experiment? Well, the output is down by 50 percent, as we expect. But the remaining 50 percent is not all x-spin $= +1$. It's half x-spin $= +1$ and half x-spin $= -1$. And the same thing happens, also contrary to our expectations, if we insert a wall in the "$x = -1$" path instead.

And now things are getting positively weird.

Consider an electron which passes through the apparatus when the wall is out. Consider the possibilities as to which route that electron could have taken. Could it have taken "$y = +1$"? Apparently not, because electrons which take *that* route (as we've just seen again) are known to have the property that their *x*-spin statistics are 50–50, whereas an electron passing through our apparatus with the wall out is known to have *x*-spin $= +1$, with certainty, at "$y = +1$ *and* $y = -1$." Could it have taken "$y = -1$," then? No, for the same reasons. Could it somehow have taken both routes? Well, suppose that when a certain electron is in the midst of passing thorough this apparatus, we stop the experiment and look to see where it is. It turns out that half the time we find it on "$y = +1$," and find nothing at all on "$y = -1$," and half the time we find it on "$y = -1$," and find nothing at all on "$y = +1$." Could it have taken neither route? Certainly not. If we wall up both routes, nothing gets through at all!

Something breathtakingly deep, it would seem, has got to give; and indeed it has become one of the central dogmas of theoretical physics over the past half-century or so that something *does*, that (more particularly) these experiments leave us no alternative but to *deny* that the very *question* of which route such an electron takes through such a contraption *makes any sense*.

The idea is that asking such questions amounts to a misapplication of language, that it amounts to something like a *category mistake*.

Thus, what typically gets said of such electrons in physics textbooks is emphatically *not* that they take either the "$y = +1$" route or the "$y = -1$" route or both routes or neither route through the apparatus, but that there is simply not any matter of fact (not merely no *known* matter of fact, but no matter of fact *at all*) about which route they take, that they are in what physicists call a *superposition* of taking the "$y = +1$" route and the "$y = -1$" route through the apparatus.

▲▲▲ Notwithstanding the profound violence all this does to our earlier picture of the world, to the very idea of what it is to be *material*, to be a *particle*, a compact set of rules has been cooked up (a set of rules which is called *quantum mechanics*) which has proven extraordinarily successful at predicting all of the thus-far-observed behaviors of electrons under the circumstances we have just been talking about, and which (as a matter of fact) has proven extraordinarily successful at predicting all of the thus-far-observed behaviors of

all physical systems under *all* circumstances, and which has functioned for more than seventy years now (as everybody knows) as the framework within which the entirety of the enterprise of theoretical physics is carried out.

Now, the mathematical object with which quantum mechanics represents both the instantaneous states and the full dynamical conditions of physical systems (which amount to *one and the same thing*—remember—for *quantum-mechanical* systems) is called the *wave-function*. Representing things that way (according to quantum mechanics) represents them *completely*, which is to say that absolutely everything that's the *case* about any given physical system at any given temporal instant (that is, the value of every single physical property of that system whose value there is—at that instant—any determinate matter of fact about, and the *probability* of any particular *outcome* of, any particular *measurement* one might choose to *carry out* on that system at that instant, *whether there is at the present instant any matter of fact about the value of the property to be measured or not*) can be *read off* (according to quantum mechanics) from its wave-function.

In the particularly simple case of a single-particle system of the sort we've been concerned with over the last few pages, the quantum-mechanical wave-function takes the form of a straightforward function of (among other things) *position* in *space*. The wave-function of a particle which is located in some spatial region A, for example, will have the value zero everywhere in space *except* in A, and will have a *non*-zero value *in* A. Similarly, the wave-function of a particle which is located in some other region *B* will have the value zero everywhere in space except in *B*, and will have a *non*-zero value *in B*. And the wave-function of a particle which is in a *superposition* of being in region A and in region *B* (the wave-function, for example, of an initially *x*-spin = +1 electron which has just passed through a *y*-box) will have non-zero values in *both* of those regions, and will be zero everywhere else.

Anyway, what the laws of physics are about, according to quantum mechanics (and indeed all the laws of physics *could* be about, *all there is* for the laws of physics to be about, according to quantum mechanics), is how the wave-functions of physical systems evolve in time. And it is an extraordinary peculiarity of the standard textbook formulation of quantum mechanics that there are two very different categories of such laws, one of which applies

when the physical systems in question are *not* being directly observed, and the *other* of which applies when they *are*.

The laws in the first category are usually written down in the form of differential equations of motion.[6] And those equations are designed to entail (for example) that an initially x-spin $= +1$ electron which is fed into a y-box will emerge from that box (just as it actually does) in a superposition of traveling along the "$y = +1$" route and traveling along the "$y = -1$" route. Moreover, all the experimental evidence we currently have suggests that those laws turn out to be the laws which govern the evolutions of the wave-functions of *all* isolated microscopic physical systems whatsoever, under *all* circumstances. And so (since microscopic physical systems are after all what everything *else* in the world *consists of*) there would seem on the face of it to be very good reason to suppose that those linear differential equations are the true equations of motion of the entire physical universe.

And yet there are reasons why (if wave-functions are indeed complete descriptions of physical systems, as quantum mechanics maintains) this can't possibly be quite right.

To begin with, the laws expressed by those equations are completely deterministic, whereas there seems to be an element of pure chance (as I discussed above) in the outcome of a measurement of (say) the position of an electron which is initially in a superposition of being in region *A* and being in region *B*. Moreover (and more important—insofar as the business of seeing precisely what it is that's *out of kilter* here is concerned), it can be shown that what the above-mentioned differential equations of the motion of a quantum-mechanical system would predict about a measuring process like that (if those equations were indeed the true equations of motion of the whole world) is emphatically *not* that the measurement would either find the electron in *A* or that it would find the electron in *B* (which is what happens when you actually go and *do* measurements like that), but rather that, with certainty, a *superposition* of those two outcomes would occur. What those equations would predict (to put it slightly differently) is that such a

6. Which is to say, the laws in this first category are usually written down in a mathematical form a good deal like the standard mathematical form of the *Newtonian* laws, except that the quantum-mechanical equations, unlike the Newtonian ones, involve (as I mentioned above) only first derivatives with respect to time.

measuring device would end up, with certainty, in a physical condition in which there is simply no matter of fact about where its pointer is pointing. It hardly needs saying, though, that this (whatever this *is*, precisely) is *not* what happens when you actually *do* such a measurement!

Thus (the standard reasoning goes) the first category of laws needs to be supplemented with a second, which will be explicitly probabilistic, and which will entail (for example) that if the position of an electron whose wave-function looks like the one in Figure 7.4 (that is, an electron about whose position there is, at present, according to quantum theory, no matter of fact; an electron which is at present in a superposition of being located in region *A* and in region *B*) were to be *measured*, then there would be a 50 percent chance of finding that electron in region *A* (which is to say, there would be a 50 percent chance of that electron's wave-function being altered, in the course of the measurement, to one whose value is zero everywhere other than in region *A*) and a 50 percent chance of finding it in region *B* (which is to say, there would be a 50 percent chance of its wave-function being altered, in the course of the measurement, to one whose value is zero everywhere other than at the point *B*).

As to the distinction between those circumstances in which the *first* category of laws applies and those in which the *second* category of laws (the laws of the so-called collapse of the wave-function) applies, all that the founders of quantum mechanics had to say was that it has something to do with the distinction between a "measurement" and an "ordinary physical process," or between what *observes* and what *is* observed, or between what lies (as it were) in *front* of measuring devices and what lies *behind* them, or between *subject* and *object*.

And it has for some time now been widely agreed to be a profoundly unsatisfactory state of affairs that the best existing formulation of the most fundamental laws of nature should depend on distinctions as imprecise and elusive as those.

And the problem of what to *do* about that, the problem of how to *fix that up* (which has emerged over the past thirty years or so as the central problem at the foundations of quantum mechanics), has gone by a number of names: the problem of Schrödinger's Cat, for example, and the problem of Wigner's Friend, and the problem of quantum state-reduction. We'll refer

Figure 7.4

to it here by its most common contemporary name, which is the *measurement* problem.

▲▲▲ There have been two big ideas (or rather, there have been two big ideas which seem to me to have any chance at all of being *on the right track*) of what to do about this problem. Both of them have been around more or less since the problem first came up; but a good deal has been learned in the past few years about what each of them really amounts to, and about how to parlay each of them into fully worked-out scientific theories.

One of those ideas (the more obvious one, given the way I've presented things here) is somehow to sharpen up the distinction between the circumstances in which the first and the second categories of laws of evolution apply; or (better yet) to cook up some *single* law of the evolutions of wave-functions which somehow *reduces*, under the appropriate circumstances, to the two sorts of laws just discussed. And the other idea involves rejecting the second category of laws altogether.

Let me talk a little about that latter one first.

The idea there is to deny that the standard way of thinking about what it means to be in a superposition is (in fact) the *right* way of thinking about it; to deny, for example, that there fails to be any determinate matter of fact, when a quantum state like the one I was just discussing obtains, about where the pointer is pointing. The idea (to come at it from a slightly different angle) is to construe quantum-mechanical wave-functions as *less than complete descriptions of the world*. The idea is that something *extra* needs to be *added* to the wave-function description, something that can broadly be thought of as *choosing between* the two conditions superposed here, something that can be thought of as somehow *marking* one of those two conditions as the unique, *actual*, outcome of the measurement that leads up to it.

And probably the most famous and probably the most successful way of

parlaying that idea into a full-fledged physical theory is due to David Bohm (and this, by the way, is precisely the theory I was talking about in footnote 8 of Chapter 4). What Bohm presented (in 1952) was in effect a *replacement* for standard quantum mechanics, which stipulates that the linear differential quantum-mechanical equations of motion are the correct equations of the time-evolutions of all quantum states at all times and under all circumstances, but on which certain facts *over and above* the facts about the quantum state of a system need to be specified in order to specify precisely and uniquely what the instantaneous physical state of that system *is*.

And what those extra facts are about are the *positions of the particles* of which the system in question is made up.

Bohm's theory presumes (notwithstanding all the evidence to the contrary presented above, of which more later) that particles *are*, after all, the sorts of things that are *invariably* located in *one or another particular place*.

Moreover, on Bohm's account, the wave-functions which are at the center of the quantum-mechanical description of the world are no longer merely (as it were) descriptive mathematical objects, but *physical* ones, physical *things*. Wave-functions, according to Bohm's theory, are somewhat like classical force-fields; and what wave-functions *do* in Bohm's theory (just as force-fields do in classical mechanics) is to sort of push the particles around, to guide them (as it were) along their proper courses.

The laws which govern the evolutions of those wave-functions in time (which, as I said, are stipulated to be precisely the same differential quantum-mechanical equations of motion discussed above, but this time with no exceptions whatsoever), and the laws which dictate how those wave-functions push their respective particles around (which are unique to Bohm's theory), *are all fully deterministic*.

And what that means, more particularly, is that the positions of all the particles in the world at any time, and the world's complete quantum-mechanical wave-function at that time (which together compose the complete instantaneous state of the world at that time, on Bohm's theory), can in principle be calculated with certainty from the positions of all the particles in the world and the world's complete quantum-mechanical wave-function at any *earlier* time; and any incapacity to carry out those calculations, any *uncertainty* in the results of those calculations, is necessarily (according to this theory) an *epistemic* uncertainty, a matter of ignorance, and not a matter of

the operations of any irreducible element of *chance* in the fundamental laws of the world.[7]

Nonetheless, this theory entails that some such ignorance (precisely enough, and of precisely the right kind, to reproduce the familiar statistical predictions of quantum mechanics by means of roughly the sort of probabilistically weighted *averaging over what one doesn't know* that goes on in classical statistical mechanics) exists for us as a matter of *principle*, some such ignorance is unavoidably forced upon us by the *laws of evolution* of the theory. The dynamics acts so as to prevent us from ever knowing enough about the physical state of the world to make those predictions which the standard irreducibly statistical formalism of quantum mechanics can't make *for* us. There is, on this account, a very real and concrete and lawlike and deterministic physical process, a process which can be followed out in exact mathematical detail, whereby the physical act of measurement unavoidably gets in the way of what is being measured. This theory entails that there is a sort of ignorance which is *merely* ignorance (merely, that is, ignorance of a certain intelligible fact about the world), and which nonetheless *could not be eliminated* without a violation of physical law, without (that is) a violation of one or another of the three fundamental laws I have mentioned over the past several paragraphs—the two dynamical ones and the one about the probabilistic averaging over one's ignorance—from which everything else about Bohm's theory follows.[8]

▲▲▲ The account that Bohm's theory produces of the experiments with the two-paths contraption (the experiments, that is, which seemed to imply

7. And note that the determinism in question here (since in Bohm's theory, as in standard quantum mechanics, instantaneous states and dynamical conditions are identical with each other) is somewhat *stronger* than it was in the Newtonian case: whereas in Newtonian mechanics the complete physical history of the world can be determined in its entirety (by means of the laws) from any one of that history's finitely long *sub-intervals*, in *Bohm*'s theory the complete physical history of the world can be determined in its entirety (by means of the laws) from any single one of the world's *instantaneous states*.

8. In *this* respect, of course, the probabilities that come up in Bohm's theory are of a very *different* kind (and an *interestingly* different kind) than the ones at the foundations of classical statistical mechanics. And that all this is so is intimately tied up (as it turns out) with the issues of the so-called compatibility of dynamical laws with non-dynamical probability-distributions that we were talking about in footnote 8 of Chapter 4. The whole story is a bit too involved to be adequately laid out here, I think, but the interested reader will have no trouble reconstructing it for herself from the materials available in (say) Chapter 5 of *Quantum Mechanics and Experience*.

that electrons can be in states in which there fails to be any matter of fact about *where they are*) runs roughly as follows.

Consider (as we did above) the case of an initially *x*-spin = +1 electron which is fed into the apparatus. On Bohm's theory, that electron will take either the "$y = +1$" route or the "$y = -1$" route, period. *Which one* of those two routes it takes will be fully determined by its initial conditions, by (more particularly) its initial wave-function and its initial position, but of course certain of the details of those conditions will prove impossible, as a matter of principle, to ascertain by measurement. Anyway, the crucial point *here* is that *whichever* route the electron happens to take, its *wave-function* will (in accordance with the linear differential equations of motion) *split up and take both*. So, in the event that the electron in question takes (say) the "$y = +1$" route, that electron will nonetheless be reunited, at the black box, with that part of its wave-function which took the "$y = -1$" route; and of course how that other part of the electron's wave-function ends up *pushing the electron around*, once the two are reunited, may well depend on whether or not it happened to (say) *run into a wall* along the "$y = -1$" route; and so that other part of the electron's wave-function can (as it were) *inform* an electron which travels through the contraption along *one* route about what's going on along the *other* one.

▲▲▲ And of course (and this is more or less the whole *point*) Bohm's theory can have nothing along the lines of a measurement problem.

Notwithstanding the fact that according to Bohm's theory the linear differential equations of motion are invariably the true equations of the time-evolution of the wave-function of the entire universe (measuring devices, observers, and all!), there are also invariably definite matters of fact about the positions of particles, and (consequently) about the positions of pointers on measuring devices, and about the positions of ink molecules in laboratory notebooks, and about the positions of ions in the brains of human observers, and (to sum it up) about the *outcomes* of *experiments*.

▲▲▲ There are a number of other responses to the measurement problem on the market nowadays—the ones I have in mind here are referred to in the literature as *modal* interpretations of quantum mechanics—which start off (just as Bohm's theory does) by stipulating that the linear dynamical equations of motion are always exactly right, and that there are certain particular

properties of physical systems (let's call them the *extra* properties of those systems) whose values are determinate even in the event that the quantum state of the world fails to be an eigenstate of the operators associated with them.

On Bohm's theory, those extra properties are the positions of particles.

On modal interpretations, things are a bit more complicated: on *those* interpretations, the identities of the extra properties can *vary* from moment to moment; and those identities *depend* on what the overall quantum state of the world is, and the particular *way* in which they depend on what that overall quantum state is (that is, the explicit *rules* whereby they depend on what that overall quantum state is) is cooked up with the aim of guaranteeing that measurements always have outcomes.

Moreover, modal interpretations (unlike Bohm's theory) aren't entirely deterministic. The evolution of the quantum state of the world is of course entirely deterministic on these interpretations (just as it is on Bohm's theory), and the *rules* whereby the identities of the extra properties depend on what the quantum state of the world is are deterministic too, but the *probabilities* associated with the various possible *values* of the extra properties, on modal theories, are *real dynamical chances*.

▲▲▲ And there are still *other* strategies that (I guess) ought to get mentioned here—and the ones I have in mind *now* can all more or less be traced directly back to the work, in the 1950s, of the late Hugh Everett—for entertaining the possibility that the dynamical equations of motion are correct but incomplete. Some of *those* strategies involve enormous multiplications of the number of *physical universes* that there are supposed to be, and others (in which I myself have had a hand) resort to odd sorts of fiddling around with the connections between the *brain* states of sentient observers and their *mental* states, and still others are tied up with very fundamental ruminations about the distinction between the first-person perspective and the third-person perspective; but the best thing (I think) will be not to get too deep into any of that right now.[9]

9. A number of these strategies are discussed in detail in Chapter 6 of *Quantum Mechanics and Experience*, and several new and novel and far more philosophically sophisticated ones (the ones about first-versus-third-person perspective) have been explored (since the time that book was published) by Simon Saunders of Oxford University. None of them (as it happens, and insofar as I can tell at present) seems to me to have a particularly good shot at turning out to be *true*—but that's a matter for some other occasion.

▲▲▲ Let me, rather, say something about the *former* of the two big ideas that I mentioned some pages back about what to do about the measurement problem, the more *obvious* one. The idea there (remember) is to stick with the standard way of thinking about what it means to be in a superposition, and to stick with the idea that a quantum-mechanical wave-function amounts, all by itself, to a complete description of a physical system, and to account for the emergence of determinate outcomes of experiments like the one we were talking about before by means of explicit *violations* of the deterministic differential equations of motion, and to try to develop some precise idea of the circumstances under which those violations occur.

There is, as I mentioned above, an enormously long and mostly pointless history of speculations in the physical literature (speculations which have notoriously hinged on distinctions between the "microscopic" and the "macroscopic," or between the "reversible" and the "irreversible," or between the "animate" and the "inanimate," or between "subject" and "object") about precisely what sorts of violations of those equations are called for here; but there has to date been only one fully worked-out, traditionally scientific sort of proposal along these lines, which is due to Giancarlo Ghirardi and Alberto Rimini and T. Weber, and which has been developed somewhat further by Philip Pearle and John Bell.

Ghirardi, Rimini, and Weber's idea (the GRW theory) goes (roughly) like this: the wave function of any single-particle system almost *always* evolves in accordance with the linear deterministic equations of motion; but every now and then (once in something like 10^9 years), at random, but with fixed probability per unit time, the wave-function is suddenly multiplied by a narrow bell-shaped curve—a curve (more particularly) whose width is something on the order of the diameter of a single atom of one of the lighter elements—which has the effect of *localizing* it, of setting its value at zero everywhere in space except within a certain small region. The *probability* of this bell curve's being centered at any particular point x depends (in accordance with a precise mathematical rule) on the *wave-function* of the particle at the moment just *prior* to that multiplication. Then, until the next such "jump," everything proceeds as before, in accordance with the deterministic differential equations.

That's the whole theory. No attempt is made to explain the occurrence of these "jumps"; that such jumps occur, and occur in precisely the way stipu-

lated above, can be thought of as a new fundamental law, a beautifully straightforward and absolutely explicit law of the so-called collapse of the wave-function, wherein there is no talk at a fundamental level of "measurements" or "recordings" or "macroscopicness" or anything like that.

Moreover, the theory can more or less do its job.

Note, to begin with, that for isolated microscopic systems (that is, systems consisting of small numbers of particles) "jumps" will be so rare as to be completely unobservable in practice.

By contrast (and this is the payoff), it turns out that the effects of these jumps on the evolutions of the wave-functions of *macroscopic* systems (systems like *measuring devices*, for example) can sometimes be dramatic. And as a matter of fact a reasonably good argument can be made to the effect that these jumps will almost *instantaneously* convert superpositions of *macroscopically* different states like {particle found in A + particle found in B} into *either* {particle found in A} *or* {particle found in B}, and that they will do so in very good accordance with the standard quantum-mechanical probabilities governing the outcomes of measurements like that.[10]

Anyway, there has been an extraordinarily lively comparative discussion going on, over the last fifteen years or so, of all the strategies for reacting to the measurement problem that I've been talking about here, and of a number of others too. And a great deal—far too much to be adequately summarized in a book which is (after all) about the foundations of *statistical* mechanics—has been learned from that discussion about the advantages and disadvantages of various of those strategies. And (nonetheless) the question of which of those strategies is the *right* one, and (moreover) the question of whether *any* of those strategies is the right one, both remain radically open. And the rest of what I want to say here can be read (I guess) as a small further contribution to all that.

10. Precisely *how* good of an argument this is, though, has been a matter of considerable (and philosophically interesting) controversy—and perhaps it ought to be mentioned that this is a question about which my own opinions have substantially evolved since I wrote *Quantum Mechanics and Experience*. What I thought back then (to make a long story short) was that the argument was pretty *bad*, and what I think now is that (as a matter of fact) it *isn't*. The details would take us much too far afield here, but they can all be very straightforwardly dug up (if the reader is interested) out of papers like "Tails of Schrödinger's Cat" (by D. Albert and B. Lower, in *Perspectives on Quantum Reality*, ed. Robert K. Clifton, Dordrecht: Kluwer Academic Publishers, 1996) and (as they say) of references therein.

What the punch line of this chapter is supposed to be—and what all of the remainder of this chapter is going to be about—is (more particularly) that the sort of story that the GRW theory has to tell about how quantum-mechanical probabilities *make their appearance in the world* (that is, as an element of *real dynamical chance*—or rather, as an element of the *right sort* of real dynamical chance, of which more in a minute—in the evolution of the world's *overall quantum state*) has in it the makings of a revolution in the foundations of statistical mechanics.

3. THE BASIC IDEA

Think of two macroscopic bodies whose temperatures initially differ.

And suppose that those two bodies are brought into thermal contact with each other. And suppose that they are not subsequently disturbed.

Over the next ten minutes, then, the temperature difference between those two bodies will decrease.

And the traditional statistical-mechanical *explanation* of that decrease, both in the classical *and* in the quantum case, runs (roughly) as follows.

The initial macrocondition of this two-body system—the one in which the two bodies are in thermal contact with each other and their temperatures are different—is compatible with a continuously infinite collection (call it {C}) of that system's possible *microconditions*. And the microconditions in {C} come in two different *varieties:* the *normal* ones (which are the ones that happen to be sitting on trajectories which pass—ten minutes hence—through a macrocondition of the two-body system in which the temperature difference between the two bodies is *lower*, and lower by the right *amount*) and the *abnormal* ones (which are all the rest, the ones associated with *un*-thermodynamic or with *anti*-thermodynamic sorts of behaviors, the ones in which the temperature difference will subsequently *rise*, or not change *at all*, or *oscillate*, or whatever). And there happens to be a breathtakingly straightforward *measure* on the set of the possible microconditions of a system like this one which is preserved by the equations of motion[11] and which our experience of the world seems to suggest is something along the lines of a measure of

11. The sort of "preservation" I have in mind here is the one connected with Liouville's theorem, or (alternatively) with the quantum-mechanical *correlate* of that theorem, which is called the principle of *unitarity*.

non-dynamical *probability*. And it happens that this measure counts the collection of *normal* points in {C} as vastly *larger* than the collection of *abnormal* points in {C}.

And that (according to the usual story) is that.

▲▲▲ But look at this.

It happens (to begin with) that the collection of normal microconditions is vastly larger than the collection of abnormal ones—on the above-mentioned standard measure—not only over the *entirety* of {C}, but over every individual not-unimaginably-small *microscopic neighborhood* of {C}, and (more particularly) over every individual not-unimaginably-small microscopic neighborhood of every individual *abnormal microcondition* of {C}, *as well!*

And what that *means* (or at any rate, *one* of the things it means) is that the property of being a *normal* microcondition is extraordinarily *stable* under small perturbations of those two bodies, and that the property of being an *abnormal* microcondition is extraordinarily *unstable* under small perturbations of those two bodies.[12]

And what *that* means is that if the two bodies we've been talking about here were in fact somehow being frequently and microscopically and randomly *perturbed, then* the temperatures of those two bodies would be overwhelmingly likely to approach each other *no matter which one* of the microconditions in {C} actually initially obtained.

The *question*, of course, is about where perturbations like that might imaginably *come from*. And the suggestion I want to make (as the reader will no doubt already have guessed) is that the *quantum jumps* in the GRW theory turn out to be just the sorts of perturbations we *need*. The suggestion (more particularly) is that it's going to turn out to be a consequence of the full stochastic dynamics of the GRW theory[13] that *every single individual one* of the microconditions in {C} will be overwhelmingly likely to evolve, over the subsequent ten minutes, into *other* microconditions in which the temperature difference between the two bodies is *smaller*, and (moreover) smaller by *precisely the right amount*.

12. This is just the thing that I made a pathetic attempt at illustrating back in Figure 3.15.

13. Which is to say, it is going to be a consequence of that dynamics *alone*; it is going to be a consequence of that dynamics without any non-dynamical addenda whatsoever.

And so if this suggestion is correct, and if anything along the lines of the GRW theory should turn out to be *true* (which, of course, is a matter for future experiments to determine) then the probabilities of universal statistical mechanics are (as a matter of fact, when you come right down to it) nothing other than the familiar probabilities of *quantum mechanics*. And if this suggestion is correct, and if anything along the lines of the GRW theory should turn out to be true, then the tendency of the temperatures of the two bodies we've been talking about here to approach each other over time amounts to a genuine (albeit statistical) *dynamical law*.[14] And if this suggestion is correct, and if anything along the lines of the GRW theory should turn out to be true, then the tendency of the temperatures of the two bodies we've been talking about here to approach each other over time can be understood entirely in terms of *readily observable characteristics* of the elementary microscopic *constituents* of those bodies—in precisely the same way that (say) the functioning of a mechanical clock can be understood entirely in terms of the material characteristics, and the spatial arrangements, of its *parts*.[15]

▲▲▲ And it happens that none of the other attempts to solve the measurement problem that have been mentioned here—and (as a matter of fact) none of the other attempts to solve the measurement problem of which I have any knowledge, and (as a matter of fact) nothing else that has ever seriously been put forward as a fundamental dynamical theory of the world—can do *anything* like that.

And all this will be worth going into in some detail, as it seems to have had a way (here and there) of uncannily escaping people's attention.

It has often been suggested in the literature (for example) that nothing even remotely as up-to-date as quantum mechanics is going to be *required* here—that (more particularly) the sorts of perturbations we were talking about above are already *all over the place*, if one simply stops and looks, in (say) the *Newtonian* picture of the world. The idea is that since none of the

14. That there simply cannot *be* any such genuinely lawlike tendency as that, on the *traditional* statistical-mechanical account of the world, was (remember) the upshot of the discussion of the pseudo-Maxwellian demon on pages 103–105 of Chapter 5.

15. Of course, the question of such an understanding, of such an *explanation*, in the case of *traditional* statistical mechanics, cannot even *arise*, since (in that case—as the reader was reminded again in the previous footnote) there is simply *not* any such lawlike tendency to be *explained*!

macroscopic two-body systems of which we have ever had any experience and none of the macroscopic two-body systems of which we ever *will* have any experience are genuinely *isolated* ones, the perturbations in question can be seen as arising simply from the interactions of the two-body system we've been talking about here with *its environment*. But if (as these authors always suppose) whatever constitutes the environment of these two bodies evolves in accord with precisely the same sorts of deterministic dynamical laws as the constituents of the bodies *themselves* do, then whatever "randomness" there is in the perturbations arising from *interactions* with that environment can only have gotten there in virtue of precisely the same sort of probability-distribution over that environment's *initial conditions* that we have been dealing with *throughout this book*. And so the whole exercise gets us nowhere.

What about something like *Bohm*'s theory? Bohm's theory has probabilities in it. The trouble is that those probabilities don't get inserted into the world *in the right place* to do the sort of job we have in mind for probabilities *here*. The only sorts of things that turn out to be *probabilistic* according to Bohm's theory (you will remember) are the *positions* of the *particles*. The only sorts of fundamental probabilities *there are* in Bohm's theory are (more particularly) probabilities that such-and-such a collection of particles has such-and-such a spatial configuration at such-and-such a temporal instant *given* that the particles' *wave-functions* have such-and-such an overall *shape* at that instant. And it happens that those parts of the laws of physics which govern the *time-evolutions* of the shapes of wave-functions, on Bohm's theory, are completely *deterministic*; and it turns out that there are wave-functions compatible with the initial macrocondition of (say) the two-body system I talked about before which (if those laws are right) will with certainty evolve, with the passage of time, into ones which determine that the temperature-difference between the two bodies will very likely have *increased*.

Modal theories have chances in them too, of course. And the chances in modal theories (unlike the ones in Bohm's theory) are genuinely *dynamical* ones. And yet the trouble here remains more or less the same: the chances aren't *in the right places*. Everything that's chancy in modal theories—just as it is in Bohm's theory—is of or about the *extra* variables. And those chances—just as in Bohm's theory—are entirely *controlled* by the *wave-functions*. And

the laws of the *evolutions* of those wave-functions (once again) are completely deterministic. And there are wave-functions compatible with the initial macrocondition of the sort of two-body system discussed before which (if those laws are right) will with certainty evolve, with the passage of time, into ones which determine that the temperature difference between the two bodies will very likely have increased.

And there are even *collapse theories*, on which the time-evolution of the wave-function *itself* is genuinely (and *dynamically*) probabilistic, which are nonetheless incapable of underwriting the foundations of statistical mechanics in the way that the GRW theory can. These sorts of theories (which have been defended in recent years by Roger Penrose, among others) stipulate that departures from the deterministic equations of motion require a "trigger"; that only certain particular *sorts* of wave-functions, the ones corresponding to superpositions of "macroscopically different states," ever undergo "collapses." And the trouble with *that* (insofar as the question of statistical mechanics is concerned) is that one can cook up (or at any rate one *fears* that one can cook up) initial wave-functions of thermodynamic systems which pick out perfectly deterministic entropy-decreasing future trajectories which entirely *avoid* those triggers.

And so the business of underwriting the thermodynamic regularities of the world, on any of the proposals for making sense of quantum mechanics I know of, with the sole exception (of course) of the GRW theory, is going to call for a story about why it is that the above-mentioned sorts of initial wave-functions—notwithstanding that they surely *exist*—need not *worry* us too much; which is to say that the business of underwriting the thermodynamic regularities of the world on any of those other theories is going to call for something along the lines of a *probability-distribution* over *initial wave-functions*, a probability-distribution which (note) is altogether *unrelated* and *in addition* to the probabilities with which those theories underwrite the statistical regularities of *quantum mechanics*.

4. THE OUTLINES OF A (POSSIBLE) NEW UNIVERSAL STATISTICAL MECHANICS

The business of deciding whether or not to take a GRW-based statistical mechanics seriously (if that turns out to be a project worth undertaking *at all*; if, that is, there should turn out to be experimental evidence that there *are* such

things in the world as collapses of wave-functions) will presumably involve detailed quantitative examinations of a host of particular cases; but there are reasons for being optimistic (and the sort of thing I have in mind here, of which more in a minute, is the very same radical instability of the condition of *abnormality* by which all of this was first *suggested*) about how those examinations will ultimately come out.

Here's the idea.

Think (to begin with) of some particular individual GRW *jump*. And call the microcondition of the system in question just *prior* to that jump A, and call the microcondition of the system in question just *after* that jump *B*.

And note that the *laws* of jumps like that (which I have already written down here, in their entirety) will straightforwardly entail an infinite set of *probability-distributions* $P_A(B)$ over all the possible *destinations* of any particular such jump, given the point at which that jump *starts out*.[16]

And there are two particular features of the $P_A(B)$'s of the GRW theory (and of any theory more or less in the *neighborhood* of the GRW theory) that it will be well (for the purposes of the next paragraph or so) to bear in mind: one is that every particular one of the $P_A(B)$'s of the GRW theory turns out to be more or less centered on its own particular A, and the other is that the volume of the space of possible microconditions over which any particular one of the $P_A(B)$'s of the GRW theory has non-negligible *values* will typically be far smaller than the volume of any one of the *macroconditions* of anything that deserves the name of a *thermodynamic system*.

Now the sort of thing we *need* from these jumps—in order to get the *statistical-mechanical* job done—is (of course) for them to be very good at getting us from *abnormal* microconditions to *normal* ones. The sort of thing we need (that is) is for it to be the case that the scales over which the tiny individual clots of abnormal microconditions typically *extend* are vastly *smaller* than the scales over which the values of the $P_A(B)$ appreciably *vary*. The sort of thing we need (more particularly) is for it to be the case that the probability of abnormality that follows from every single individual one of the $P_A(B)$'s (no matter what A may happen to be) is roughly equal to the probability of abnormality that follows from the standard statistical-mechanical measure

16. And indeed, the complete set of those probability-distributions, together with the probability per unit time of a jump's *occurring*, is the entirety of what those laws have to *say*.

over the *entirety* of the *macrocondition* within which the *A* in question happens to *fall*.[17]

And it would seem to be an eminently plausible proposition—given the radical unimaginable submicroscopic *tinyness* of the *clots*, and given the two particular characteristics of the jumps in the GRW theory that we took note of in the paragraph before last—that there are any *number* of different sorts of GRW-like perturbations that are perfectly capable of getting all that accomplished.

▲▲▲ Nonetheless, there are *hard* cases, or *apparently* hard ones (and I am thankful to Larry Sklar and Phillip Pearle, among others, for bringing these to my attention); and there turn out to be interesting lessons in them; and it will be worth taking the trouble to think through two or three of them in some detail.

Consider (for example) an extraordinarily tiny gas, one which consists of something on the order of 10^5 molecules. Even gasses as tiny as that are known to be very likely to *spread out* (if space is available) over reasonable intervals of time, and yet gasses as tiny as that are very *unlikely* to suffer even a single GRW-type *collapse* over such an interval, and so an explanation of the tendencies of gasses like that to evolve like that over intervals like that in terms of GRW-type collapses of the wave-functions of their constituents is apparently out of the question.

Or consider the collection of dazzling and beautiful experiments which have actually been performed over the past twenty years or so, and which are

17. Actually, we don't need quite that, and probably can't quite *have* it. The trouble (and here we will have no alternative but to help ourselves rather freely to some of the technical jargon of quantum theory) is that abnormal quantum states have got to be more or less *orthogonal* (if you think about it) to *normal* ones, and that no single GRW collapse can ever (in and of itself) bring about transitions between states that *are* (perfectly) orthogonal to one another, and that (as a matter of fact) no single GRW collapse is ever going to be able to do much of *anything* (in and of itself) about the abnormality of a quantum state if that state should happen to be anything along the lines of an eigenstate of the *positions* of the particles that make the system in question up. But none of that turns out to matter much. Let the *A* and *B* we have just now been discussing represent (instead) the before and after states of a dynamical process involving (say) *two* GRW collapses, or *three*, or *twenty*, with the appropriate deterministic dynamical evolutions *between* them (all of which is still going to be overwhelmingly likely to take place, on the GRW theory, over time-intervals which are negligibly short compared—say—with times over which the temperatures of the two bodies we were talking about before ever undergo any significant change)—and everything will come out fine.

referred to in the scientific literature as "spin-echo" experiments, in which it has turned out to be possible to isolate some very large array of interacting microscopic systems from the relevant sorts of external influences—and (moreover) to replace the *dynamical condition* of that array, at a certain particular instant, as the array is in the midst of some entropy-increasing transformation, with its *time-reverse*—and (thereafter) merely to watch, in astonishment, as the array traces its previous trajectory out, dutifully, *backward*.

The microscopic systems in question are typically atomic nuclei. And these nuclei are typically being held at fixed spatial positions—but in such a way that the orientations of their nuclear magnetic fields are free to rotate—by intermolecular forces in a crystal. And the sort of thing that happens in these experiments is (very schematically) that the nuclei are all initially arranged with their magnetic fields pointing in the same direction—and then they're left (as it were) to their own devices, and they magnetically interact with one another, and their magnetic fields begin to pivot around, and in time the directions in which those individual fields are pointing become more and more disorganized and uncorrelated. Eventually a state of *equilibrium* is arrived at, in which the arrangement of the individual fields is *random*, in which (that is) the cumulative *macroscopic* magnetic field of the entire array is *zero*, and then (and this is the cool part) a very intense *external* magnetic field is turned on for a very short time, which has the effect (for reasons that need not concern us here) of turning all those tiny individual fields exactly *around*—and then the system is left again to its own devices, and in time, and (more particularly) in precisely the same *amount* of time as had elapsed between the array's first having been left to its own devices and the moment when the external field was turned on, the fields spontaneously *realign* themselves![18]

It would seem (on the face of it) that GRW collapses can play no role

18. These spin-echo systems are (by the way) fairly close relatives of the pseudo-Maxwellian demons we were talking about in Chapter 5. Both of them (like the *bona fide* demons) are in direct violation of the letter of the second law. And both of them (*unlike* the bona fide demons) systematically falsify the predictions of the uniform-over-the-present macrocondition probability-distribution about the future evolution of the world. And neither of them (unlike the bona fide demons) produces any net increase, at the end of the day, in the proportion of the total energy of the universe which is available for routine mechanical exploitation. But note that whereas the pseudo-demons we were thinking about in Chapter 5 need to be able to ascertain certain microscopic details of present conditions of the systems on which they operate, nothing of that sort needs doing in the spin-echo case.

whatsoever in any *explanation* of the initial approach to equilibrium here. The trouble is that the atomic nuclei in these experiments are very rigidly *held in place*—which is to say that the *wave-functions* of those atomic nuclei are permanently *localized*—which is to say that the wave-functions of those atomic nuclei are permanently frozen into that particular mathematical form which is (if you think about it) altogether *impervious* to the effects of GRW collapses—by the powerful intermolecular forces I mentioned above. Moreover (and this is the *particularly* astonishing business—and this seems powerfully confirmatory of the doubts expressed in the previous two sentences), it turns out that the approach to equilibrium can be *reversed*—it turns out that the original alignment of the fields can be *reinstated*—simply by flipping the nuclei around!

Or consider what it is, on a statistical mechanics of the sort that we have been imagining here, that guarantees that a regular-sized gas in equilibrium at t will not spontaneously explode or condense or turn into an elephant between t and $t + \&$, where $\&$ is so short an interval that even a regular-sized gas is unlikely to suffer a GRW-type collapse in it.

Let's think through these three cases one at a time.

Take the case of a small gas. We might appeal, there, to the fact that we have no empirical experience whatsoever, that (come to think of it) we *can* have no empirical experience whatsoever, of a small gas which is *genuinely isolated from all external influences*. And so for all we now empirically know or ever *will* empirically know, it might not *be* a law of nature that gasses like that tend to spread out *at all*! And the behaviors of the sorts of small gasses that can actually be *looked* at can very plausibly be accounted for by GRW-type collapses of the wave-functions of particles in (say) their *containers*.

Or we could appeal to the fact that such gasses, even if they *are* isolated, have *pasts*. This will take a bit more setting up. What we will want to show, in this case, is that the GRW theory will entail that a small isolated gas which is condensed at t, and which is around for a while, is likely to be more dispersed at $t + @$, even if the gas in question is unlikely to undergo a single collapse in the interval between t and $t + @$. Good. Here's how to do it: call the average time between GRW collapses in the gas in question i, and call the gas's macrocondition at t C, and call the gas's macrocondition at (say) $t - i(100000)$ S. And consider the probability, on the GRW theory, given that the macrocondition at $t - i(100000)$ is S and that the macrocondition at

t is C, that the *microcondition* of the gas at *t* will be one of the "normal" ones. And note that the instability of the property of being abnormal will entail, completely independent of what state *S* is, that that probability is high.

What about the case of the spin-echo experiments? Collapses in the environment will patently get us *nowhere* with *that*. The realignability of the fields, after all, amounts to a direct empirical *proof* that those collapses (just like the ones that hit the nuclei themselves) produce no significant short-term disruptions of the trajectories along which this system evolves. But the longer term is (of course) another matter. Given sufficient time, even in systems like *this*, GRW collapses will move us relentlessly away from abnormality. And so there would seem to be every reason in the world to believe that the *previous history* of the array of nuclei in question, *whatever* that history may have been, will give us just what we need.

What about large gasses over the very *short* term? The environment will be of no avail there either; but histories still will. And here a *third* strategy suggests itself. The macroconditions of thermodynamic systems never get measured at *instants*. The thermodynamical regularities of our actual experience, if you stop and think about it, are relations between the physical situations of systems not *at* different instants but *around* different instants. And so maybe the right way to think of propositions like "this is a gas with such-and-such a volume and a temperature and a pressure" is to see them as asserting that certain physical properties of a certain collection of particles have *persisted* over a certain short interval. And if we read such propositions *that* way, they will entail (in conjunction with the GRW theory) that the probability that the microcondition of the gas in question is a normal one is high.

You get the idea. The crux of the matter is that the job of statistical mechanics is *not* (after all) to underwrite the *letter* of the laws of thermodynamics, but to underwrite the actual content of our thermodynamic *experience*. And I know of no compelling argument, at present, why a statistical mechanics based on GRW collapses should be incapable of doing that.

▲▲▲ One can go further. If the GRW theory should turn out to be *true*—and this, of course, is a very big *if*—it may turn out (as I mentioned earlier on) that there is at bottom only a single kind of probability in nature. It may turn out (that is) that all the robust lawlike statistical regularities there are,

not only in thermodynamics but (one can even imagine) in biology, and in psychology, and in sociology, and God knows where else, are at bottom nothing other than the probabilities of certain particular GRW collapses' hitting certain particular sub-atomic particles.

▲▲▲ As to the question of Maxwellian demons, they are plainly going to be more or less as much in accord with the laws of physics, and they are going to be more or less as difficult to actually *construct*, in the context of the sort of fundamental theory of the world we have been playing around with here as they were in the context of Newtonian statistical mechanics. All the main arguments about those demons in Chapter 5 (as the reader can easily confirm for herself) have a relatively straightforward translation into quantum-mechanical language, and (thence) into the language of the GRW theory.

The only *exceptions* (insofar as I can see) are going to arise in cases where there are relatively long intervals over which the evolution of the Newtonian version of the system in question somehow substantially *departs* from what the uniform-over-the-current-macrocondition-probability-distribution and the deterministic equations of motion jointly predict—systems (say) like the *pseudo*-Maxwellian demons we were talking about in Chapter 5, the ones which rearrange the microconditions of boxes of gas in such a way that at some particular *later* time, the gasses in those boxes (which are then evolving as fully isolated systems) will spontaneously begin to *contract*.

▲▲▲ And (finally) as to the question of the overall logical structure of the universal GRW-statistical-mechanical contraption for making inferences, it comes out like this:

There are *two* fundamental laws (as opposed to the three in the standard contraption) and one contingent empirical fact.

The empirical fact is (as before) the one about what the macrocondition of the world currently happens to *be*, and the laws are:

1. The GRW law of motion—the GRW law (that is) of the time-evolutions of quantum-mechanical *wave-functions*.
2. The *past-hypothesis* (which is, again, that the world first came into being in whatever particular low-entropy sort of macrocondition it is

that the normal inferential procedures of cosmology will eventually present to us).

And that's it.

A few remarks are in order.

To begin with (and this is more or less the whole point of the exercise), this contraption contains nothing whatsoever along the lines of a *statistical postulate*. All the statistics there *are* in this theory (which is to say, all the statistics there are in any *world* of which this theory turns out to be the correct fundamental scientific description) are the purely *quantum-mechanical* ones in the GRW equations of motion. This is an account of the world into which chance enters exactly *once*, and (as I've been saying) there seems to me to be no other known strategy for making sense of quantum mechanics on which anything like that can possibly be true. On modal theories, for example, there are going to be dynamical chances in the fundamental microscopic laws of motion—just as there are here—but (since the chances in the modal theories turn out not to be *situated in the world* in such a way as to be able to get the statistical-mechanical job done) those chances are going to need to be *supplemented*, in the context of any universal statistical mechanics, with a non-dynamical *statistical postulate* (very much along the lines of the one cooked up for *Newtonian* mechanics) which stipulates some probability-distribution over initial universal *wave-functions*. And much the same sort of thing is going to be necessary on *Bohm*'s theory, and on collapse theories with triggers in them, and on everything else I know of. All of them (that is) are going to need to adopt precisely the sort of universal statistical contraption we worked out for Newtonian mechanics back in Chapter 4, in which (now) two utterly unrelated sorts of chance are going to appear—one (the *quantum-mechanical* one) in the fundamental microscopic equations of motion, and the other (the *statistical-mechanical* one) in the statistical postulate.

And note that the *past-hypothesis* is playing a slightly different conceptual *role* here than it did in the contraption at the end of Chapter 4. The reason a past-hypothesis needed to be added to the contraption back in Chapter 4 (remember) was that *without* such a hypothesis, the contraption turned out to generate all sorts of claims about the past which were radically *false*. And the reason that hypothesis needed to be written in such a way as to refer to *the*

very first instant of the existence of the world was that if it were written in any *other* way, if it were written in such a way as to refer to any *other* past instant, then the full contraption (*including* that past-hypothesis) would generate all sorts of claims about times *prior* to that past instant which are radically false. And the situation in a GRW-based universal statistical mechanics is going to be altogether different. The GRW contraption, minus the past-hypothesis, makes *no claims whatsoever* (statistical or otherwise) about the past. And so the necessity of a past-hypothesis arises here not as a matter of *correcting an error*, but (as it were) as a matter of filling a space which is cleanly and transparently *left empty* for it by the mathematical structure (and more particularly by the *time-reversal asymmetry*)[19] of the microscopic equations of motion. And the reason that hypothesis *now* needs to be written in such a way as to refer to the very first instant of the existence of the world is that if it were written in any *other* way, if it were written in such a way as to refer to any *other* past instant, *then* the full contraption (including that past-hypothesis) would fail to generate *any claims whatsoever* about times prior to that past instant, and there would still be space left (as it were) to fill.

19. Note (for example) that no such space is left empty here for anything along the lines of a *future*-hypothesis.

APPENDIX / INDEX

APPENDIX
GEDANKENEXPERIMENTS WITH HEAT ENGINES

I demonstrated in the text that Kelvin's formulation of the second law is deducible from Clausius's formulation of it, together with one or two auxiliary stipulations—which happen to be empirically true—to the effect that certain particular thermodynamic transformations are possible.

And I mentioned that (given our empirical knowledge of the possibility of certain *other* transformations) *Clausius*'s formulation of the second law is also deducible from *Kelvin*'s. And the first thing I want to do here is to show how that deduction goes.

The transformations involved in this case are called *Carnot cycles*. And saying precisely what those are will take some setting up. Consider two bodies at different, uniform temperatures $t1$ and $t2$. And suppose (just to keep things simple and neat) that these bodies are large enough that significant quantities of heat can be removed from them or added to them without significantly *affecting* those temperatures. And suppose there is a much smaller gas, in a container with a piston, which is initially at temperature $t2$ (the higher one), and which is initially in thermal contact with the *large* body at $t2$. And suppose that the following sequence of events takes place: (1) While the gas is in thermal contact with the body at $t2$, the piston is slowly, reversibly, pulled out a certain distance.[1] (2) The gas is thermally isolated, and the piston is reversibly pulled out a bit farther, until the temperature is reduced to $t1$. (3) The gas is put into thermal contact with the body at $t1$, and the piston is reversibly *pushed in* to the point where (4) a final ther-

1. If the piston is pulled out slowly enough, the temperature of the gas will remain constant throughout this process: whatever energy the gas loses to work on the piston it immediately reabsorbs as heat from the large body at $t2$.

mally insulated reversible compression will return the gas precisely to its original volume, temperature, and pressure. Insofar as the gas is concerned, then, all this amounts to a *cycle*.

Let's get in deeper. Part of what emerged from the discussions of the billiard balls in the text was that a gas does *work* in pushing a piston out, and that it has work done *on* it when a piston pushes it *in*. And a little further reflection on those discussions will show that the *amount* of work involved, in both cases, is proportional to the product of the pressure of the gas and the change in its volume. And it turns out to follow from *this* that the net work done by the gas on the external world throughout the course of the cycle described above is positive, and is proportional to the area enclosed by the path the cycle traces out in a pressure-volume diagram. And so the first law of thermodynamics will require that the heat absorbed by this gas at $t2$ exceeds the heat it relinquishes at $t1$.

And the *name* of the sort of contraption I have been describing here is a *heat engine*. In typical heat engines—in *steam*-engines, for example—the higher-temperature body (the body from which heat is *extracted*) is something like a *boiler*, and the lower-temperature body (the body into which heat is *relinquished*) is something like the *atmosphere*; and the sort of *mechanical energy* such engines produce is typically something like *the twisting of a crank shaft*. And note that it follows from the fact that these engines convert heat into work, and from the fact that they operate in *cycles*, that they must necessarily operate between two *different* temperatures. Absorbing heat from the boiler, and converting it entirely into work, and relinquishing none of it into the atmosphere, and leaving the thermodynamic state of the world otherwise unchanged, would amount (after all) to a direct violation of Kelvin's formulation of the second law.

Anyway, given the possibility of cycles like these, Clausius's formulation of the second law can be deduced (as promised) from Kelvin's. It goes like this: suppose (contradicting Clausius's formulation) that a certain quantity of heat *could* be made to flow, without any other thermodynamic changes in the world, from (say) the body at $t1$ to the body at $t2$. Then, running a Carnot cycle between the two bodies, one which removes a *larger* quantity of heat from the body at $t2$, turns (as above) some of that heat into work, and dumps the rest (which is arranged so as to be equal to the quantity that flowed from the cooler body to the hotter one at the outset) into the body at $t1$ (leaving

the body at $t1$ in precisely its original state), would complete a process in which heat is removed from the body at $t2$ and converted into work with no other net changes in the thermodynamic state of the world. And this would of course amount to a violation of Kelvin's formulation of the second law.

▲▲▲ Let's go further.

The *efficiency* with which a cyclic heat engine converts heat into work is defined as the amount of work produced per cycle of the engine divided by the amount of heat extracted, per cycle, from the higher-temperature body. And it turns out that the possibility of executing *Carnot* cycles (or of very *nearly* executing them, at any rate),[2] combined with either the Clausius or the Kelvin formulation of the second law, allows the calculation a specific quantitative upper limit on the possible efficiency with which any cyclic heat engine whatsoever, operating between two specified temperatures, can function.

The first step will be to prove that no cyclic engine operating between two specified temperatures $t1$ and $t2$ can possibly have a higher efficiency than a *reversible* engine operating between those temperatures.[3] Here's how to do that: consider two cyclic engines, one of which is reversible, operating between $t1$ and $t2$. Let the reversible one be operating in reverse, as a work-*consuming* refrigerator. Arrange things (and this may involve running both engines repeatedly, and different numbers of times) so that at the end of a full combined cycle the $Q2$'s of these two engines (one of which is positive and the other of which is negative) will exactly cancel each other. Then, if the *irreversible* engine had the higher efficiency, the combined device (since the magnitudes of the two $Q2$'s are equal) will have a positive net work output—in a situation where heat is removed from body 1 and no net thermodynamic changes occur in any *other* systems—in violation of Kelvin's formulation of the second law.

And note that if *both* engines are reversible, it will follow from the above argument that their efficiencies must necessarily be *equal*.[4]

2. There are, after all, no such things in the world as *perfectly* reversible thermodynamic transformations. Pistons, for example, are never utterly free of *friction*. But the friction can presumably be made as small as one likes, if one is willing to take the trouble.

3. Note that a heat engine can perfectly well be *cyclic* without being *reversible*. *Carnot* engines, of course, are both.

4. If both engines are reversible, after all, *either one* can be the one running backward in the above argument.

And so (since Carnot cycles are examples of reversible cyclic engines, and since they can actually be *instantiated*, and their efficiencies *tested*) the upper limit on the efficiency of any cyclic engine operating between two given temperatures $t1$ and $t2$, which is equal to the *actual* efficiency of any *reversible* cyclic engine operating between those two temperatures, can be quantitatively *known*. Moreover, fixing the efficiency of a heat engine amounts to fixing the value of the ratio $Q1/Q2$—the efficiency, by definition, is (the work output)/$Q2 = (Q2 - Q1)/Q2 = 1 - (Q1/Q2)$—and *that* number turns out to be equal, for Carnot engines, to $T1/T2$.

And the interest of all this is that it will facilitate the formulation of a *vastly* more informative version of the second law—one that will stipulate, *quantitatively*, *what* changes in the rest of the world are required in order to *pull off* the transformations mentioned (and forbidden, on their own) in the Clausius and Kelvin formulations.

Here's how to get at that: imagine that a certain system undergoes a cyclic transformation in the course of which it absorbs positive or negative amounts of heat (call these amounts Qk) from a number of bodies (whose temperatures are Tk). And let there be a number of reversible cyclic heat engines around, operating between these various bodies and one *other*, at temperature $T0$; and let them operate so as to restore the original heat contents of all those bodies (except the one at $T0$). Now it will follow from what we learned above about the efficiencies of reversible heat engines that the total amount of heat surrendered by the body at $T0$ in the course of this restoration process is $Q0 = T0$(the sum over k of $[Qk/Tk]$). And since this process in entirely cyclic except insofar as the body at $T0$ is concerned, whatever heat is absorbed from that body has necessarily been transformed entirely into *work*, and so, in order to avoid contradiction with Kelvin's formulation of the second law, that amount ($Q0$) had better not be positive, which means that the sum over k of $[Qk/Tk]$ had better not be positive as well. Moreover, if the original cyclic process is reversible, $Q0$ (and hence also the sum over k of $[Qk/Tk]$) must obviously be *zero*.

This can now easily be parlayed into a demonstration that the sum over k of $[Qk/Tk]$ for any *reversible* route from one thermo-state A to another thermo-state B must be *equal* to the sum over k of $[Qk/Tk]$ for any *other* reversible route between those two states. The argument is just that any two reversible routes between the same initial and final states can always be con-

verted (by running one of them forward and then running the other one backward) into a thoroughly reversible *cycle*—and we have just shown that the sum over k of $[Qk/Tk]$ over any such cycle is zero. And so the sum over k of $[Qk/Tk]$ over any reversible route between two distinct thermodynamic states is (as was promised in the text) a function only of the initial and final thermo-states in question and is referred to (as mentioned) as the *entropy*-difference between them. And thus if we pick, by convention, a zero-entropy state for a given sort of system, the entropy becomes a unique and definite *function* of the thermodynamic state—a new, bona fide thermodynamic variable (whose *status* as a variable is guaranteed by the second law of thermodynamics).

Moreover, the above considerations *also* entail that the Q/T sum over *irreversible* routes between two specified thermodynamic states (call them A and B) must be *less than* or equal to the value of the sum over reversible ones—which is to say that the integral over any irreversible route between A and B must be less than or equal to the *entropy-difference* between A and B.

Thus, if B can emerge out of A by means of a transformation which is both reversible and *isolated* (that is, Q-exchange = 0), then the entropies of A and B must be identical; and conversely, if the only isolated transformations by means of which B can emerge out of A are *irreversible* ones, *then* the entropy of state B must be *greater than or equal to* the entropy state A.

And so (and this is the punch line—this is the now-canonical formulation of the second law) the total entropy of the world (or of any isolated subsystem of the world), in the course of any transformation, either keeps the same value or goes up.

▲▲▲ Note, by the way, that we now know something quite definite, something *quantitative*, about the "thermodynamic changes in the rest of the world" mentioned in the *Clausius* formulation of the second law. We now know (as a definite function of the two temperatures involved) *how much more* heat needs to be dumped into the hotter body by any cyclic refrigerator than that refrigerator *removes* from the *cooler* one.

INDEX

This book is an attempt to get to the bottom of an acute and perennial tension between our best scientific pictures of the fundamental physical structure of the world and our everyday empirical experience of it. The trouble is about the direction of time. The situation (very briefly) is that it is a consequence of almost every one of those fundamental scientific pictures—and that it is at the same time radically at odds with our common sense—that whatever can happen can just as naturally happen backwards.

"Albert is perfecting a style of foundational analysis that is uniquely his own . . . It has a surgical precision . . . and it is ruthless with pretensions. The foundations of thermodynamics is a topic that has accumulated a good deal of dead wood; this is a fire that will burn and burn."

SIMON W. SAUNDERS
Oxford University

"As usual with Albert's work, the exposition is brisk and to the point, and exceptionally clear . . . The book will be an extremely valuable contribution to the literature on the subject of philosophical issues in thermodynamics and statistical mechanics, a literature which has been thin on the ground but is now growing as it deserves to."

LAWRENCE SKLAR
University of Michigan

DAVID Z ALBERT is Professor of Philosophy at Columbia University and author of *Quantum Mechanics and Experience* (Harvard).

HARVARD UNIVERSITY PRESS
Cambridge, Massachusetts, and London, England
www.hup.harvard.edu

PAINTING: Fortunato Depero, *Train Born from the Sun*, 1924. © 2000 Artists Rights Society (ARS), New York / SIAE, Rome.
DESIGN: Jill Breitbarth

DAVID Z ALBERT TIME AND CHANCE