THE CONCEPT OF IRREVERSIBILITY IN STATISTICAL MECHANICS

ROBERT ZWANZIG

Institute for Fluid Dynamics and Applied Mathematics and Institute for Molecular Physics, University of Maryland

ABSTRACT

The interaction between concepts of irreversibility and the development of non-equilibrium statistical mechanics is discussed with particular reference to certain 'paradoxes' that retarded this development. A recently evolved attitude towards the general problem of irreversibility may be considered responsible for a number of practical advances in this field.

My subject is the interaction between concepts of irreversibility and the development of non-equilibrium statistical mechanics. In particular, I will discuss certain 'paradoxes' that seriously retarded this development, and how these paradoxes are viewed in modern work. Then I will describe an attitude towards the general problem of irreversibility, evolved in the last ten or fifteen years, that in my opinion is responsible for a number of important practical advances in non-equilibrium statistical mechanics.

Everyone knows what irreversibility is. On a primitive level, we know that fire burns wood to ashes, that men grow old and die, and that taxes will increase.

On a more advanced level, we know that irreversibility is associated with an inevitable increase of entropy. We know from experience that we are unable to construct devices to decrease the total entropy of our local environment. Whatever we are able to do, the entropy increases. Our *experience* is summarized in the second law of thermodynamics.

On a still more advanced level, the irreversible increase of entropy is given a cosmic generality and a deep philosophical significance. Here, the standard view is summarized in the famous statement by Clausius: '. . . *die Entropie der Welt strebt einem Maximum zu*'. In this form, the second law of thermodynamics is often used to discuss the approach of the entire universe to a state of thermal equilibrium, or to support our intuition that the flow of time has an 'arrow' attached to it.

Many of the difficulties that arise in the statistical mechanical theory of irreversibility can be traced to the sweeping generality of the third view of irreversibility that we have just referred to. At this level, there seems to be a fundamental contradiction between the second law and mechanics. The main point I want to make here is that these difficulties can be avoided by taking a more modest point of view, in which the second law merely summarizes certain human experiences. I do not say that one *must* take the more modest view, but I say that *if* one takes this view, *then* one can develop a practical non-equilibrium statistical mechanics.

My story begins in 1872 with Boltzmann and the kinetic theory of gases. Boltzmann had two concerns. One was a practical one, to be able to predict various transport properties of gases, for example their viscosity coefficients. The other concern was to understand the mechanical basis for the thermodynamic concept of an equilibrium state. He introduced the Boltzmann equation, which is an integro-differential equation for the evolution of the kinetic distribution function $\mathrm{f}(v; t)$, the density of gas atoms having velocity v at time t. In the course of his investigations, he discovered that a particular functional of the velocity distribution,

$$H(t) = \int \mathrm{d}v \, \mathrm{f} \log \mathrm{f}$$

had an extremely interesting time dependence. On computing the evolution of $H(t)$ with the Boltzmann equation, he found that it can never increase with time, or

$$(\mathrm{d}H/\mathrm{d}t) \leqslant 0 \,(\text{all } t)$$

It can only decrease or remain constant; and it remains constant only in the state of thermodynamic equilibrium.

Thus $H(t)$ shows the same kind of irreversible behaviour that we expect of the entropy. And, in fact, H is the negative of the entropy for an equilibrium state.

It should be emphasized that Boltzmann's H-theorem is an exact consequence of the Boltzmann equation for $\mathrm{f}(v; t)$. Further, we know (at least in retrospect) that the Boltzmann equation provides a correct and useful theory of simple transport processes in dilute gases.

But almost immediately, certain objections were raised to the validity of the H-theorem and the Boltzmann equation. These objections, in various forms, plagued subsequent investigations in non-equilibrium statistical mechanics for many years.

The first objection, usually called the 'reversibility paradox', was raised by Lord Kelvin and by Loschmidt. In modern terms, this paradox may be called a violation of time-reversal symmetry. The fundamental equations of motion of any conservative dynamical system, e.g. a monatomic gas, are Newton's, Lagrange's or Hamilton's equations. These equations are invariant to the substitution of $-t$ for t; or, they are symmetric to time reversal. The H-theorem and the Boltzmann equation violate this symmetry, so they cannot be consistent with any exact dynamical theory. Therefore they cannot be correct.

The second objection to Boltzmann's work was raised by Zermelo and by Poincaré, and is usually called the 'recurrence paradox'. It arises whenever one deals with a finite closed dynamical system.

If a system of interacting particles is confined to a closed region of space, and if their interaction energy has a finite lower bound, then the motion of the system is confined to a finite region of phase space. According to ergodic theory, the trajectory of the system in phase space passes arbitrarily closely to any assigned position on the surface of constant total energy; given sufficient time, it does so arbitrarily often. So any given state of the system will recur to within any assigned accuracy. This indicates that a gas contained in a finite volume cannot approach an equilibrium state and then stay there indefinitely. Any non-equilibrium state that was passed through once will be visited again if one waits long enough.

(In quantum mechanics, the same objection takes the following form. If the system is enclosed in a finite region of space, its energy level spectrum must be discrete. Then the time dependence of any dynamical property is given by an almost-periodic function, and recurrences are guaranteed.)

The objections that we have quoted, and a number of variations on them, led to the general impression that no statistical mechanical theory based on exact dynamics could be consistent with the irreversibility that we observe in nature.

Because of this, workers in the field felt that one must 'do something' to the exact equations of motion before irreversibility would emerge. Very often, lectures on new methods in non-equilibrium statistical mechanics involved heated discussion of the question 'Where did you put the irreversibility into the theory?' Many ingenious answers were given.

One approach that has been popular for a long time is called 'coarse graining'. This has both classical and quantum mechanical forms, but for illustration I use the classical one. It is argued that the exact position of any system in phase space is never observed experimentally and is of no interest. One should divide phase space into cells of finite size, each cell corresponding roughly to some macroscopic description of the state of the system; and one should focus attention not on the detailed distribution within any individual cell, but only on the net content of that cell.

A variant of this view is called 'time-smoothing'. Here it is argued that we are unable to observe experimentally the precise time dependence of any dynamical quantity; because of inherent limitations on our apparatus, only a time average is observed. The time interval used for averaging is supposed to be short compared with characteristic times of macroscopic processes, but long compared with characteristic times of elementary molecular processes. Sometimes time-smoothing is used along with coarse graining.

Objections can be made to these ideas. One objection is that smoothing techniques are based on the assumed limitations of experimental apparatus. We know from experience, however, that measurement techniques are being constantly refined, and are more and more delicate. Any theory based on assumed limitations of this kind is likely to be superseded some day.

Another objection should be mentioned. Generally speaking, smoothing techniques are only productive if also certain *discarding* operations are performed. Somehow, information has to be thrown away. A careful inspection of theories based on these techniques will disclose that smoothing is always accompanied by various approximations that have the effect of discarding information. In my opinion, these approximations (sometimes quite subtle ones) are vital to the success of smoothing techniques. The operation of smoothing is itself only a convenient way of leading up to the necessary approximations.

Another procedure used to 'put irreversibility into the theory' is to suppose that the system is not contained in a closed box. This is justified by the observation that any real system must interact with its surroundings, and they must interact with *their* surroundings, *ad infinitum*.

In some instances, this method for introducing irreversibility is quite efficient. Consider, for example, the spontaneous emission of a photon from an excited atom. If the photon is able to escape from the atom into infinite space, then it will never return to be re-absorbed, and the process is irreversible. If the accessible space is finite, then eventually the photon must be reflected by a wall, so that it can return to the atom and be re-absorbed.

A difficulty with the open system approach is that the only true equilibrium state is that of the entire universe; this is not much help for our own petty concerns.

Another variation is to suppose that the system is contained in a box having walls that are not fixed, but move randomly according to some stochastic process. This procedure will also lead to irreversibility. While it works, however, the stochastically moving walls may often be regarded as irrelevant. A clock that is going to run down and stop will do so whether or not it is located in some perfectly sealed room, and whether or not the walls of the room are jiggling. In a closed system, we expect that the clock will eventually start up again, but this is of no interest to anyone who wants to know what time it is during his own lifetime.

During this long period of investigation into the foundations of non-equilibrium statistical mechanics, significant progress was made with respect to more practical questions. It soon became clear that the Boltzmann equation gave a valid, experimentally verifiable description of transport processes in gases at low enough densities. The theory of Brownian motion was developed. In the early days of quantum mechanics, the theory of transition rates between quantum states (as expressed in the 'golden rule') was worked out. Onsager derived his reciprocal relations and showed how one could make practical use of non-equilibrium thermodynamics.

All these advances in useful technique for handling non-equilibrium systems were made without regard to the fundamental difficulties connected with the 'paradoxes' I just discussed.

As attempts were made to extend the limits of validity of familiar methods, for example to derive a generalization of the Boltzmann equation that would be valid for dense gases or to construct a theory of transport processes in liquids, the impression grew that progress could be made only after the paradoxes were resolved. This is partly the reason why so much attention was given to methods for 'putting irreversibility into the theory'.

But in the 1950s, a striking change in direction and attitude occurred. Van Hove presented a sound derivation of the Pauli master equation for weakly interacting systems, and gave a correct explanation of the validity of the theory of transition rates. Soon after that, he presented the first generalization of the master equation to strongly interacting systems.

At about the same time, diagrammatic techniques were developed for handling perturbation and other expansions to infinite order. Using these techniques, Prigogine and many others made significant progress in non-equilibrium statistical mechanics.

Also at about the same time, Kubo presented his remarkably simple analysis of the response of a many body system to an external field, leading to the time-correlation function expressions for transport coefficients. Kubo's work had a particularly strong impact, perhaps because his approach was so direct and intuitively obvious.

Another important development of the last decade was the introduction and analysis of simple analytically tractable models for non-equilibrium systems. I mention especially work by Rubin, Montroll, Mazur and others on harmonic oscillator models of Brownian motion. The importance of such simple model systems is that they lead to mathematically exact results, and can be used to test methods based on mathematical approximations.

In my opinion, however, the most important development in recent years was a change in point of view. If one looks for some feature common to recent successful theories, one finds that they all have a strongly operational character.

Consider for example Kubo's analysis of linear transport processes. Suppose that we want to find the electrical conductivity of a metal. In Kubo's theory, we construct a canonical ensemble distribution function for a piece of metal at some temperature. To the Hamiltonian describing the metal we add a perturbation term due to the interaction of the metal with an external electric field. We use perturbation theory to find out how the original equilibrium distribution function is modified by a time dependent electric field. Then we use the modified distribution function to compute the average electric current in the presence of the field. It turns out that the current is proportional to the field. Then the coefficient of proportionality must be the electrical conductivity of the metal.

Notice that each step in this procedure corresponds to an operation that one would perform in a laboratory experiment. Selection of a sample piece of metal at some temperature corresponds to starting out with an initial canonical ensemble distribution function. Switching on an external electric field corresponds to adding a perturbation to the Hamiltonian. Measuring the current corresponds to calculating the ensemble average of the current. The measured electrical conductivity is the coefficient of proportionality between the measured current and the applied field.

The success of this procedure is connected with the following statement of belief. *If each step in a statistical mechanical calculation can be put into one-to-one correspondence with a step in some experiment, then the result of the calculation must be the same as the result of the experiment.*

There are, of course, still some serious difficulties to be faced. The expression for electrical conductivity that one obtains this way may be not at all easy to calculate; the time dependence of a fluctuating electric current must be found, and an equilibrium average has to be calculated. But these are well-defined computations, not involving deep questions of principle. This is where simple model systems are useful.

Essentially the same procedure is followed in modern derivations of the quantum mechanical master equation, describing the evolution of diagonal elements of a density matrix. Here we start out with some non-equilibrium system in which the density matrix is initially diagonal; we use operator methods to follow the time dependence of the density matrix at later times; we focus attention on only the diagonal part of the density matrix at later times; and then we do what are essentially just algebraic manipulations to find a generalized master equation. Explicit computation of the coefficients in the master equation is still a hard job, but no questions of principle are involved.

Similar analyses may be carried out for many familiar transport theories, but a detailed description of all of these would be out of place here. However, one can draw a general conclusion. Successful treatments all seem to be based on the same idea—an initial state is defined carefully, dynamical processes are then followed exactly (though sometimes only in a purely formal way), and finally, only certain specific questions are asked about the results.

But we are still left with the nagging question of the paradoxes. How is it that we are able to proceed at all, in view of the asserted contradiction between exact dynamical principles and the second law? Perhaps this question is best answered by considering a simple model for which exact results can be obtained.

This is a model of Brownian motion in one dimension. Let me first remind you of the standard theory of Brownian motion. I will use the Langevin form of the theory. The Brownian particle has a mass M and a velocity $v(t)$ at time t. Its equation of motion is the Langevin equation

$$M(\mathrm{d}v(t)/\mathrm{d}t) = -\zeta v(t) + F(t)$$

where $-\zeta v(t)$ is the frictional force on the particle and ζ is the friction coefficient. The extra force $F(t)$ is a fluctuating force due to interactions of the particle with its environment, and is known only in a statistical way. In particular, it is treated as a gaussian random variable, with zero mean value, and a second moment

$$\langle F(t)\,F(t')\rangle = 2\zeta k_B T \delta(t - t')$$

T is the temperature of the medium and k_B is the Boltzmann constant. The preceding equation is often called a Nyquist formula or a fluctuation–dissipation theorem.

This standard Brownian motion theory leads to the paradoxes. Consider, for example, the time dependence of the average velocity. It is evident that $\langle v(t)\rangle$ decays exponentially from some initial value

$$\langle v(t)\rangle = v(0)\exp\left(-\zeta t/M\right)$$

It does not recur, as it should; and time reversal symmetry is violated.

Now we want to compare this standard theory with a modern one, based on a simple model in which dynamical calculations can be performed exactly. The model is due to R. J. Rubin. [His most recent publication on the subject is in *J. Am. Chem. Soc.* **90**, 3061 (1968). This gives references to earlier work.]

Rubin's model is a finite one-dimensional nearest neighbour harmonic crystal, consisting of $2N + 1$ particles with periodic boundary conditions. All particles except the one labelled 0 have a mass m, while particle 0 has a mass M which is much larger than m. The Hamiltonian is

$$H = \frac{1}{2M}p_0^2 + \sum_{\substack{j=-N\\ \neq 0}}^{N} \frac{1}{2m}p_j^2 + \frac{1}{2}K \sum_{j=-N}^{N} (r_j - r_{j+1})^2$$

p_j and r_j denote the momentum and displacement of the jth particle, and K is a force constant.

Because the model contains only coupled oscillators, the equations of motion are all linear and can be solved by matrix methods. In particular, the perturbation due to the heavy particle can be handled exactly.

Now let us construct an experiment. At the initial time $t = 0$, the heavy particle has a given momentum $p_0(0)$. All other momenta, and all displacements, are assumed to have a thermal equilibrium distribution at temperature T. Thus the initial state is well defined in a statistical mechanical sense. What then is the motion of the heavy particle?

At time t, $p_0(t)$ can be expressed as a linear combination of terms, each of which is a product of a known function of time and some initial momentum or displacement. This is a consequence of the linearity of the equations of motion. By working backwards, one can find a generalized Langevin equation for $p_0(t)$

$$(\mathrm{d}/\mathrm{d}t)\,p_0(t) = -\int_0^t \mathrm{d}s\,k(s)\,p_0(t-s) + F(t)$$

In this, $k(s)$ is a time dependent friction coefficient, and $F(t)$ is a fluctuating force. Further, $F(t)$ is a gaussian random variable. (It is a linear combination of initial values of all displacements and all momenta except p_0, and these are supposed to have a Boltzmann distribution.) The mean value of $F(t)$ vanishes, and its second moment is given by a generalized fluctuation–dissipation theorem

$$\langle F(t)\,F(t')\rangle = Mk_BTk(t-t')$$

Notice that the only difference between this theory and the standard Langevin theory is in the time dependence of the function $k(s)$. If this were a delta-function, then the standard theory would be recovered. (The missing factor of two comes from taking one half the delta-function in the convolution over time.)

Rubin succeeded in calculating analytically the quantities needed to obtain $k(s)$; to avoid excessive detail, I will not write the final result here. More important, he was able to solve analytically for the time dependence of the *average* momentum of the heavy particle. This makes it possible to compare the predictions of the exact calculation with those obtained from the standard Langevin theory.

Without going into detail, I will describe in qualitative terms what the results are. First, the average momentum is an even function of time. This is a consequence of the exact character of the calculation. Time reversal symmetry is not violated. Secondly, if the calculations are performed for a finite lattice, then recurrences in the average momentum are found. This again is a consequence of the exact character of the calculation. We must conclude that no paradoxes are evident in this model.

The third and most important property of the average momentum is found when both the mass ratio M/m and the lattice size N are very large. Then the decay of the average momentum is *approximately exponential.* The precise meaning of this statement was worked out by Rubin.

Three significant time scales are observed. The first is a short one, of the order of the reciprocal of the Debye frequency of the uniform lattice,

$$t_1 = \tfrac{1}{2}(m/K)^{\frac{1}{2}}$$

During this interval, the average momentum is not exponential; rather, it is approximately parabolic in time, with a maximum at $t = 0$. But the change in magnitude of the average momentum is only of order m/M during this interval.

The next time scale is much longer

$$t_2 = (M/m)t_1$$

For times of this order of magnitude, approximate exponential decay is found, with the relaxation time t_2. The difference between the exact result and the exponential approximation is small. Rubin gave a numerical estimate of the orders of magnitude involved when the mass ratio is $M/m = 10^4$. He found that '... after 18 relaxation times, the correction to the exponential is less than 10^{-4} of the value of the exponential'. It seems to me that it would be extraordinarily difficult to detect such a small deviation from exponential decay in any real experiment.

The third time scale is defined by

$$t_3 = Nt_1$$

For times of this order of magnitude, recurrences will be seen. If we had done the calculation for an infinite lattice in the first place, this time scale would never appear; but we would still see the same exponential decay. If the lattice is finite and large enough, we would never see recurrences in our own lifetimes.

I emphasize that this theoretical model of Brownian motion does not involve any smoothing process, stochastic element, or any other act of violence on exact dynamics. Irreversibility is not 'put into' the theory anywhere.

So here we have an answer to our question about the apparent contradiction between exact dynamical principles and the second law. As long as we are willing to settle for a theory of irreversibility on a human, experimentally observable time scale, there are no contradictions. If we confine our efforts in non-equilibrium statistical mechanics to problems that are rooted in operationally well-defined experiments, and if we have the courage to do hard calculations, then we are bound to succeed.