1 The Direction of Time *

Oliver Penrose[1]

Heriot-Watt University, Riccarton, Edinburgh EH14 4AS, UK

To appear in *Chance in Physics : Foundations and Perspectives,* edited by J. Bricmont, D. Dürr, M. C. Galavotti, G. C. Ghirardi, F. Petruccione and N. Zanghi (SpringerVerlag 2001)

Abstract. It is argued, using a relativistic space-time view of the Universe, that Reichenbach's "principle of the common cause" provides a good basis for understanding the time direction of a variety of time-asymmetric physical processes. Most of the mathematical formulation is based on a probabilistic model using classical mechanics, but the extension to quantum mechanics is also considered.

1.1 Does time have a direction?

When we speak of "the direction of time" we are not really saying that time itself has a direction, any more than we would say that space itself has a direction. What we are saying is that the events and processes that take place in time have a direction (i.e. they are not symmetrical under time reversal) and, moreover, that this time direction is the same for all the events.

To see the distinction, imagine that you look into a body of water and see some fish all facing in the same (spatial) direction. You would not attribute the directedness of this fish population to any "direction of space"; it would be much more natural, assuming the body of water to be a river, to attribute the directedness of the fish population to the direction of a physical phenomenon which is external to them: that is, to the direction of the flow of the river.

An external influence of this kind is not the only mechanism by which a set of objects (or creatures) can line up along a particular direction: they may line up because of an interaction between them rather than some external influence. It might be that the water is stagnant, imposing no direction on the fish from outside, but that the fish happen to like facing in the same direction as their neighbours. In this case there is an interaction between neighbouring fish, but the interaction itself is invariant under space reversal: if two fish are facing the same way and you turn both of them around, they should be just as happy about each other as before. In such cases, when an asymmetrical arrangement arises out of a symmetrical interaction, we speak of spontaneous symmetry breaking.

* This paper is dedicated to the memory of Dennis Sciama (1926-1999), the sadly missed friend who taught me so much about cosmology, the foundations of physics, and scientific writing.

In the case of the time asymmetry, we are surrounded by examples of processes that are directed in time, in the sense that the time-reversed version of the same process would be virtually impossible (imagine raindrops rising through the air and attaching themselves to clouds, for example); moreover, as we shall see in more detail later, the time direction of each process is the same every time it happens (imagine a world where raindrops fell on some planets but rose on others[1]). Just as in the case of the asymmetrical spatial arrangement of fishes, there are two possible explanations of such a large-scale asymmetry: either it is due to some overarching asymmetry which separately causes the time direction of each of the individual processes, or else it is due to some linkage between different events which, though itself symmetric under time reversal, leads most or all of them to adopt the same time direction as their neighbours so that an asymmetry can arise through spontaneous symmetry breaking.

The most obvious physical linkage between different processes comes from the fact that the same material particles may participate in different processes at different times; the states of the same piece of matter at different times are linked by dynamical laws, such as Newton's laws of motion which control the motion of the particles composing that matter. Processes involving different particles can also be linked, because of interactions between the particles, and the interactions are subject to these same dynamical laws. The dynamical laws are symmetrical under time reversal; nevertheless, in analogy with our example of the fishes, this symmetry is perfectly consistent with the asymmetry of the arrangement of events and processes that actually takes place.

In this article I will look at and characterize the time directions of various types of asymmetric processes that we see in the world around us and investigate whether all these asymmetries can be traced back to a single asymmmetric cause or principle, or whether, on the other hand, they should be attributed to some form of symmetry breaking. The conclusion will be that there is, indeed, a single asymmetric principle which accounts for the observed asymmetry of most physical processes. Many of the ideas used come from an earlier article written with I. C. Percival[17], but others are new, for example the extension to quantum mechanics.

1.2 Some time-directed processes

Let's begin by listing some physical processes that are manifestly asymmetrical under time reversal.

The subjective direction of time. Our own subjective experience gives us a very clear distinction between the future and the past. We remember the past, not the future. Through our memory of events that have happened to us in the past, we know what those events were; whereas a person who "knew" the future in such a way would be credited with supernatural powers. Although some future events can be predicted with near certainty (e.g. that the Sun will rise tomorrow),

[1] This possibility may not be quite so fanciful as it sounds. See [20]

we do not know them in the same sense as we know the past until they have actually happened. And many future events are the subject of great uncertainty: future scientific discoveries, for example. We just do not know in advance what the new discoveries will be. That is what makes science so fascinating – and so dangerous.

Part of our ignorance about the future is ignorance about what we ourselves will decide to do in the future. It is this ignorance that leads to the sense of "free will". Suppose you are lucky enough to be offered a new job. At first, you do not know whether or not you will decide to accept it. You feel free to choose between accepting and declining. Later on, after you have decided, the uncertainty is gone. Because your decision is now in the past, you know by remembering what your decision was and the reasons for it. You no longer have any sense of free will in relation to that particular decision. The (temporary) sense of free will which you had about the decision before you made it arose from your ignorance of the future at that time. As soon as the ignorance disappeared, so did the sense of free will.

It is not uncommon to talk about the "flow" of time, as if time were a moving river in which you have to swim, like a fish, in order to stay where you are. Equivalently one could look at this relative motion from the opposite point of view, the fish swimming past water that remains stationary; in the latter picture there is no flow as such, but instead a sequence of present instants which follow one another in sequence. The time direction of this sequence does not come from any "flow", but rather comes from the asymmetry of memory mentioned earlier. At each instant you can remember the instants in the past that have already happened, including the ones that happened only a moment ago, but you cannot remember or know the ones in the future. The time asymmetry of the "flow" is a direct consequence of the asymmetry of memory; calling it a "flow" does not tell us anything about the direction of time that we do not already know from the time asymmetry of memory.

Recording devices. Why is it, then, that we can only remember the past, not the future? It is helpful to think of devices which we understand better than the human memory but which do a similar job although they are nothing like as complex and wonderful. I have in mind recording devices such as a camera or the "memory" of a computer. These devices can only record the past, not the future, and our memories are time-asymmetric in just the same way.

Irreversible processes in materials. Various macroscopic processes in materials have a well-defined time direction attached. Friction is a familiar example: it always slows down a moving body, never accelerates it. The internal friction of a liquid, known as viscosity, similarly has a definite time direction. Other examples are heat conduction (the heat always goes from the hotter place to the colder, never the other way) and diffusion. The time directions of all the processes in this category are encapsulated in the Second Law of Thermodynamics, which tells us how to define, for macroscopic systems that are not too far from equilibrium, a quantity called entropy which has the property that the entropy of an isolated system is bound to increase rather than to decrease. The restriction to systems that are not too far from equilibrium can be lifted in some cases, notably gases, for which Boltzmann's H-theorem[3] provides a ready-made non-decreasing quantity which can be identifed with the entropy.

Radio transmission. When the electrons in a radio antenna move back and forth, they emit expanding spherical waves. Mathematically, these waves are described by the retarded solutions of Maxwell's equations for the electromagnetic field (i.e. the solutions obtained using retarded potentials). Maxwell's equations also have a different type of solution, the so-called advanced solutions, which describe contracting spherical waves; but such contracting waves would be observed only under very special conditions. Although Maxwell's equations are invariant under time reversal, the time inverse of a physically plausible expanding-wave solution is a physically implausible contracting-wave solution.

The expanding universe. Astronomical observation tells us that the Universe is expanding. Is the fact that our Universe is expanding rather than contracting connected with the other time asymmetries mentioned above, or is it just an accident of the particular stage in cosmological evolution we happen to be living in?

Black holes. General Relativity theory predicted the possibility of black holes, a certain type of singularity in the solution of Einstein's equations for space-time, which swallows up everything that comes near to it. The time inverse of a black hole is a white hole, an object that would be spewing forth matter and/or radiation. It is believed[18] that black holes do occur in our Universe, but that white holes do not.

1.3 The dynamical laws

The astonishing thing about the processes listed above is that, although they are all manifestly asymmetrical under time reversal, every one of them takes place in a system governed, at the microscopic level, by a dynamical law which is *symmetrical* under time reversal. For many of these processes, a perfectly adequate dynamical model is a system of interacting particles governed by Newton's laws of motion – or, if we want to use more up-do-date physics, by the Schrödinger equation for such a system. In the case of the radio antenna, the dynamical model is provided by an electromagnetic field governed by Maxwell's equations, and in the cosmological examples it is a space-time manifold governed by Einstein's equations. All these dynamical laws are invariant under time reversal.

To provide a convenient way of discussing the laws of dynamics and properties of these laws such as invariance under time reversal, I'll use a purely classical model of the Universe. The possibility of generalizing the model to quantum mechanics will be discussed in section 1.7. The model assumes a finite speed of light, which no moving particle and indeed no causal influence of any kind can surpass, and is therefore compatible with relativity theory.

Imagine a space-time map in which the trajectories of all the particles (atoms, molecules, etc.) in the model universe are shown in microscopic detail. In principle the electromagnetic field should also be represented. The space-time in which the map is drawn will be denoted by U. Fig. 1 shows a one-dimensional map of this kind, containing just one trajectory. For generality, we shall use (special) relativistic mechanics, with a finite speed of light, so that the space in Fig. 1 represents Minkowski space-time. (Similar diagrams could be drawn for non-flat cosmological models such as the Einstein-de Sitter model[21].)

For each space-time point X, we define its *future zone* to comprise all the space-time points (including X istelf) that can be reached from X by travelling no faster than light. A signal or causal influence that starts from X cannot reach any space-time point outside the future zone of X.

The *past zone* of X comprises those space-time points (including X itself) from which X can be reached by travelling no faster than light. Signals starting from space-time points outside the past zone of X cannot influence what happens at X. Since real particles travel slower than light, any particle trajectory passing through X stays inside the past and future zones of X.

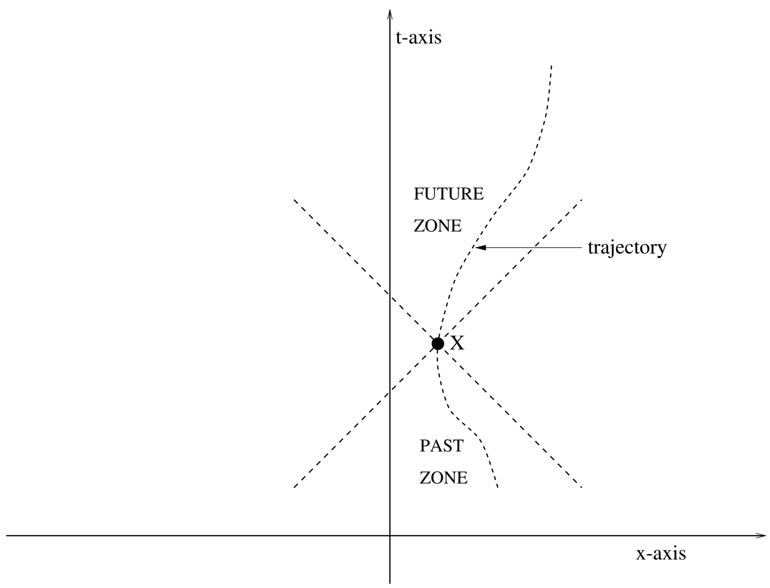

Fig. 1. Typical particle trajectory in a universe with one space dimension. The future and past zones of the space-time point X on the trajectory are shown as quadrants. The two dotted half-lines bounding the future (past) zone of X are the trajectories of the two possible light rays starting (ending) at X; their equations are $x = ct$ and $x = -ct$ where c is the speed of light – which, for the purpose of drawing the diagram, has been taken to be 1 unit.

By the *history* of the model universe being considered, I mean (disregarding the electromagnetic field for simplicity) the totality of all the trajectories in it; since each trajectory is a set of points in U, the history is also a certain set of points in U. This history will be denoted by the symbol ω. Given a space-time region A within U, we shall use the term the *history of A* to mean the intersection of A with the history of U; that is to say, the totality of all pieces of trajectory that lie inside A. The history of A will be denoted by $\omega \cap A$ or just a. See Fig. 2.

The laws of dynamics place certain restrictions on the histories; for example, if the model universe were to consist entirely of non-interacting particles, then all the trajectories would be straight lines. We shall not assume that all the trajectories are straight lines, but we shall make two specific assumptions, both of which are satisfied by Newton's equations for the mechanics of a system of particles and by Maxwell's equations for the electromagnetic field: these assumptions are that the dynamical laws are (i) invariant under time reversal and (ii) deterministic. (We could also require relativistic invariance but this is not essential). By time-reversal invariance we mean that the time inverse of any possible history (obtained by inverting the space-time map of that history) is also a possible history – or, more formally, that if R denotes the time-reversal transformation in some inertial frame of reference, implemented by reversing the time co-ordinate in that frame of reference, and if ω is any history that is compatible with the laws of dynamics, then $R\omega$ is also compatible with the laws of dynamics.

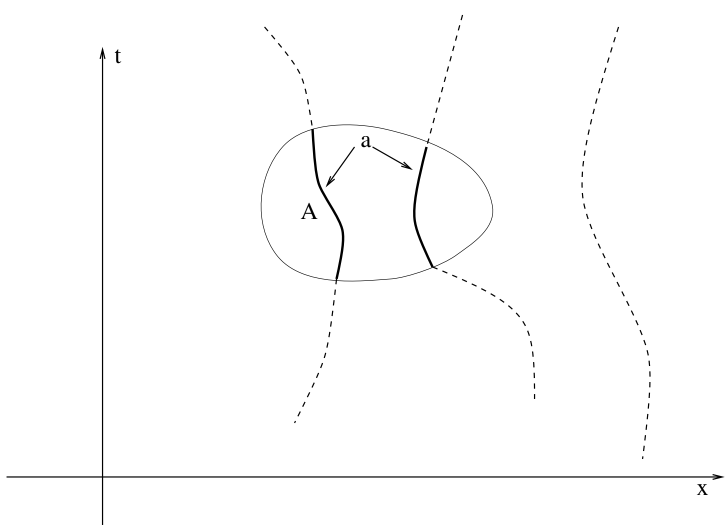

Fig. 2. The history a of the egg-shaped space-time region A consists of the the two part-trajectories which lie inside A.

The following definitions will help us to formulate the concept of determinism in this theory. By the *past zone* of a region A, denoted $PZ(A)$, I mean the union of the past zones of all the points in A. The region A istelf is part of $PZ(A)$. If A and B are two disjoint regions, it may happen that the part of $PZ(A)$ lying outside B consists of two (or more) disconnected parts, as in Fig. 3. In that case, we shall say that B *severs* $PZ(A)$. In a similar way, one region can sever the future zone of another.

Now we can formulate what we mean by determinism in this theory. If a region B severs the past zone of a region A, then any particle or causal influence affecting the history of A must pass through B. The history of B includes all these causal influences, and so it determines the history of A. In symbols, this means that there exists a function f_{AB} from the history space of B to that of A such that the laws of dynamics require

$$a = f_{AB}(b). \tag{1.1}$$

For deterministic dynamics, such an equation will hold for every pair of regions A, B such that B severs the past zone of A.

Since we are assuming the laws of dynamics to be time-symmetric, the time inverse of (1.1) also holds. That is to say, if there is a region B' which severs the *future* zone of A then the dynamical state of A is completely determined by that of B'. In other words, there exists a function $g_{AB'}$ such that the laws of dynamics requre

$$a = g_{AB'}(b'), \tag{1.2}$$

and such an equation holds for every pair of regions such that B' severs the future zone of A.

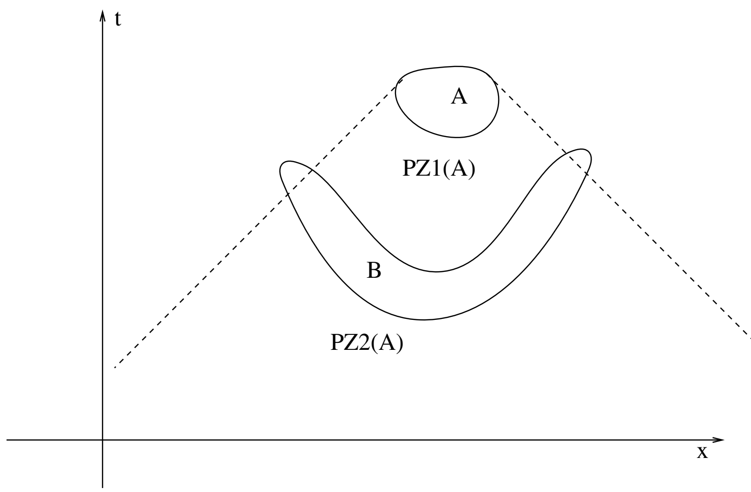

Fig. 3. Example where B severs the past zone of A. The past zone of A consists of three separate parts: one part inside B, and two disconnected parts outside B, marked $PZ1(A)$ and $PZ2(A)$. By definition, A itelf is a part of $PZ1(A)$

1.4 Probabilities: the mathematical model

As we have seen, time direction is not a property of the laws of motion themselves. The laws of motion are too general to solve the problem of time asymmetry, since they do not distinguish between a physically reasonable motion and its physically unreasonable time inverse. To characterize the time asymmetry we need to say something about which solutions of the dynamical equations are likely to occur in the real world. This can be done using probability concepts. The mathematical model I shall use will be based on the following postulates:

(i) there is a probability measure on the space Ω consisting of of all conceivable microsopic histories of the model universe U (including histories that do not obey the dynamical laws);

(ii) the measure is invariant and ergodic under spatial shifts (translations);

(iii) the actual history of U was selected at random using this measure, so that any property which holds with probability 1 under the measure (and whose definition does not mention the actual history) is a property of the actual history.

To keep the mathematical formulation of postulate (i) as simple as possible, let us assume that Ω is discrete, and that the set Ω_A of conceivable histories of any given (finite) space-time region A is finite. (The extension to the more realistic case of history spaces that are not discrete involves only standard methods of probability theory.) For each A, there is a probability distribution over the histories in Ω_A, that is to say a set of non-negative numbers $p_A(a)$, the probabilities, such that $\sum_{a \in \Omega_A} p_A(a) = 1$. The probability distributions for different regions must satisfy certain consistency relations, arising out of the fact that if one region is a subset of another then the probability distribution for the smaller region is completely determined by that for the larger; it will not be necessary to give these consistency relations explicitly.

There is no contradiction here between determinism and the use of probabilities: we are dealing with a probability distribution over a space of deterministic histories. But determinism does impose some conditions on the probability distribution, the most obvious of which is that any trajectory in Ω_A violating the dynamical laws (e.g. Newton's equations of motion) has probability zero. Determinism also imposes conditions relating the probability distributions in different regions; thus from formula (1.1) it follows that if B severs the past zone of A then there is a function f_{AB} such that

$$p_A(a) = \sum_{b \in \Omega_B} \delta(a, f_{AB}(b)) p_B(b) \tag{1.3}$$

where $\delta(\cdot,\cdot)$ is defined to be 1 if its two arguments are the same and 0 if they are different.

To formulate the first part of postulate (ii) mathematically, we assume that there is a group of spatial shifts[2] T such that if A is any region in the universe U then TA is also a region in U, and that if a lies in Ω_A (i.e. if it is a possible history of A) then its translate Ta is a possible history of TA. If A is a union

[2] Unlike the dynamical model, our probability model cannot be relativistically invariant, since the group of spatial shifts is not the full symmetry group of (special) relativistic space-time. But the lower symmetry of the probability model should not be reckoned as a disadvantage, since the whole purpose of the probability model is to identify deviations from this full symmetry, in particular deviations from time-reversal symetry.

of disjoint parts, then TA is obtained by applying T to every part. Then the postulate of shift invariance can be written

$$p_{TA}(Ta) = p_A(a) \quad (a \in \Omega_A) \tag{1.4}$$

holding for all space-time regions A and all spatial shifts T.

To a very good approximation, we would expect probabilities to be invariant not only under space shifts but also under time shifts, provided that the size of the time shift is not too big. Invariance under shifts of a few seconds, days, years or even millennia is fine, but to postulate invariance under arbitrarily large time shifts, comparable with the age of the universe or greater, would be to commit oneself to the steady-state cosmological theory, which is no longer in fashion.

The second part of postulate (ii), ergodicity, means that a law of large numbers holds with respect to space shifts: given any event E that can occur in a space-time region A, the fraction of the space translates of A in which the (spatially shifted) event E actually occurs is almost surely equal to the probability of E. In symbols, if E is a subset of Ω_A (for example E could consist of just one history) then the following statement is true with probability 1:

$$\lim_{n\to\infty} \frac{1}{n} \sum_{i=1}^{n} \chi_n(\omega) = p_A(E) \tag{1.5}$$

where χ_n denotes the indicator function for the occurrence of the event T^nE in the region T^nA; i.e. $\chi_n(\omega)$ is defined to be 1 if $\omega \cap T^nA \in T^nE$ and to be 0 if not.

By postulate (iii) eqn (1.5), being true with probability 1, is true in the actual Universe. It is this property that connects the mathematical probability model with observable properties of the real world, enabling us to equate the probabilities in it with frequencies of real events. In practice, it is easier to use time shifts than spatial ones, reproducing the same situation at different times in the same place rather than in different places at the same time. Since the time shifts used in estimating the average on the left side of (1.5) by this method are likely to be very small compared to the age of the Universe, the use of time shifts when formula (1.5) is used to estimate probabilities should be a very good approximation.

As an example, A could be any space-time region consisting of a cube with side 1 metre lasting for 1 second, starting now, and E could denote the event in that the cube is empty of particles throughout its lifetime Then $p_A(a)$ is the probability that a 1 metre cube, randomly chosen at this moment somewhere in the Universe, is empty of particles during the second after the moment when it is chosen. Knowing something about the density and temperature of the of gas in outer space, one could give a reasonable numerical estimate of this probability.

1.5 Physical probabilities

The probabilities we usually deal with in science refer to events not in outer space but here on Earth. In the present formalism, these are conditional probabilities, conditioned on some particular experimental set-up. For example, suppose you throw a spinning coin into the air and note which side is up when it lands. The probability of the outcome 'heads' can be written as a conditional probability:

$$\text{Prob}(\mathcal{H}) = p_{_{A\cup B}}(\mathcal{H}|\mathcal{B}) = \frac{p_{_{A\cup B}}(\mathcal{H} \times \mathcal{B})}{p_{_B}(\mathcal{B})} \tag{1.6}$$

where $\mathcal{H}$ represents the 'heads' outcome and $\mathcal{B}$ represents the launching of the coin and the other requirements that make the tossing of a coin possible, for example the presence of a floor and a gravitational field. The left side of 1.6 is to be thought of as a physical property of the coin, measurable by replicating the experimental condition $\mathcal{B}$ many times and counting the fraction of occasisions when the outcome is $\mathcal{H}$.

In our probability model, the macroscopic event $\mathcal{H}$ is represented by a set of histories for a space-time region A which includes the place and time where the coin comes to rest on the floor, and $\mathcal{B}$ is represented by a set of histories for an earlier space-time region B which is disjoint from A and includes the place and time where the coin is launched. The notation $\mathcal{H} \times \mathcal{B}$ denotes the set of histories $\varpi \in \Omega_{_{AB}}$ such that $(\varpi \cap A) \in \mathcal{H}$ and $(\varpi \cap B) \in \mathcal{B}$. Note that B comes earlier than A, not later: the coin is spun before the time when it is observed on the floor, not after. Moreover, to ensure that all the influences that might influence the motion of the coin are properly controlled, B should sever the past zone of A, as in Fig. 3.

By itself, eqn (1.6) is just a definition; it contains no information about the behaviour of real coins. We can put some empirical information into it, however if we take account of something that every probability theorist (though not every gambler) believes to be true, namely that there are many features of the macroscopic state $\mathcal{A}$ of the region A upon which $\text{Prob}(\mathcal{H})$ depends very slightly, if at all. The sex of the experimenter, for example, makes no difference to the probability of $\mathcal{H}$; nor does it matter (so the probability theorists believe) how many times the 'heads' outcome occurred on the previous occasions when the coin was spun. It is this independence of irrelevant features of the macroscopic state of A (a feature that is given the name "statistical stability" in [14]) that makes it possible to think of $\text{Prob}(\mathcal{H})$ as a physical quantity, which can be measured in many different laboratories to give the same answer.

Our experience of the statistical stability of physical probabilities indicates that, if we define the macroscopic states in the right way, the probabilities of macroscopic events should be independent of what happened before the region B got into the macroscopic state $\mathcal{B}$. In symbols, we expect the equation

$$p_{_{A\cup B}}(\mathcal{H}|\mathcal{B}) = p_{_{A\cup B\cup C}}(\mathcal{H}|\mathcal{B} \times \mathcal{C}) \tag{1.7}$$

to hold for all macroscopic events $\mathcal{C}$ in the history space of an extra space-time region C which is arbitrary except that it must sever the past zone of B, as shown in Fig. 4. Thus, if C is any space-time region that contains the experimenter just before the coin is spun, $\mathcal{C}$ could be the event "the experimenter is a woman"; or if C is a space-time region containing all the previous occasions when the coin was spun, $\mathcal{C}$ could be the event "all the previous spins gave the result 'heads'".

In the mathematical theory of probability, equation (1.7) is part of the definition of a Markov chain; so (1.7) can be regarded as saying that it is possible to choose the macroscopic states such as $\mathcal{B}$ in such a way that their probabilities have a Markovian structure. In [14], such a Markovian structure is taken as one of the main postulates in a deductive treatment of the foundations of statistical mechanics, and it is shown to lead to an equation expressing transition probabilities such as $p_{A\cup B}(\mathcal{H}|\mathcal{B})$ in terms of purely dynamical quantities. That discussion does not use space-time maps, however, being geared to the standard methods of statistical mechanics where the system we are interested in is considered to be isolated from the rest of the world during the period when its time evolution is studied. The approach described here arose in part from trying to avoid this restriction.

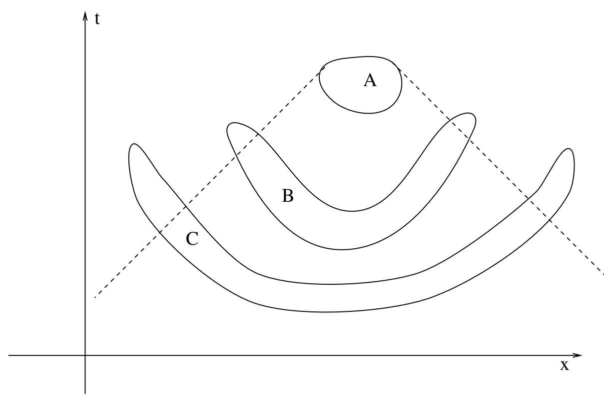

Fig. 4.The arrangement of regions A, B, C for equation (1.7). The region B severs the past zone of A and C severs that of B.

What about the time direction of physical probabilities? At first sight, equation (1.6) appears to contain a very clear time asymmetry, since in the experiments done to measure physical probability we always prepare the system before we observe it, not after: the region B in (1.6) always comes before A, not after it. But is this really a time asymmetry of the equation? Using the definition of conditional probability, and replacing $\mathcal{H}$ by a more general macroscopic event $\mathcal{A}$ in A, (1.7) can be rewritten as

$$\frac{p_{A\cup B}(\mathcal{A}\times\mathcal{B})}{p_B(\mathcal{B})} = \frac{p_{A\cup B\cup C}(\mathcal{A}\times\mathcal{B}\times\mathcal{C})}{p_{B\cup C}(\mathcal{B}\times\mathcal{C})} \tag{1.8}$$

which can be rearranged to give

$$p_{B\cup C}(\mathcal{C}|\mathcal{B}) = p_{A\cup B\cup C}(\mathcal{C}|\mathcal{B}\times\mathcal{A}). \tag{1.9}$$

Thus the conditional probability of the earlier event $\mathcal{C}$ given the later one $\mathcal{B}$ is independent of the even later event $\mathcal{A}$. The Markovian condition (1.7), despite its asymmetrical formal appearance, is in fact symmetrical under time reversal as far as the time order of the three events $\mathcal{A}, \mathcal{B}, \mathcal{C}$ is concerned. One might be tempted to conclude from the reversed Markovian condition (1.9) that the conditional probability of an earlier macroscopic event given a later one has a similar type of statistical regularity to that of the later event given the earlier one, in which case we could use the left side of (1.9) to define a "time-reversed physical probability" for the earlier event given the later one.

Nevertheless, such a conclusion would be wrong. The "time-reversed physical probability" corresponding to the left side of (1.6), namely $p_{\mathcal{B}\cup\mathcal{A}}(\mathcal{B}|\mathcal{H})$, would be the probability, given that a coin is found on the floor with its 'heads' side uppermost, that the coin arrived there as the result of a coin-tossing experiment (rather than, for example, as the result of somebody's dropping it on the floor by mistake). Such probabilities are well-defined in the model, but physicists cannot measure them without going outside their laboratories, nor philosophers without going outside their studies; they need outside information to tell them, for example, how often people actually do drop coins on the floor by mistake. In short, the time-inverse probability $p_{B\cup A}(\mathcal{B}|\mathcal{H})$ cannot be measured by laboratory experiments and therefore, unlike the left side of (1.7), is not a physically measurable property of the coin.

The reason for the wrong conclusion discussed in the preceding paragraph is that the Markovian condition resides not only in the formula (1.7) (with $\mathcal{H}$ now replaced by a general macroscopic event $\mathcal{A}$) but also in the geometrical relation between the space-time regions A, B, C appearing in it. Unlike the formula itself this geometrical relation, illustrated in Fig. 4, is not symmetrical under the symmetry operation of reversing time and interchanging the labels A and C: thus, C severs the past zone of A, but A does not sever the future zone of C.

So the Markovian condition (1.7), if true for suitably specified macroscopic events, does after all give us a direction of time, its asymmetry under time reversal deriving not from any algebraic property of the formula, but from the geometrical asymmetry of the relation between the space-time regions A and B illustrated in Fig. 3: if we reversed this relation, making B sever the future zone of A instead of the past, and C sever the future zone of B, then there would be no reason to expect either the formula (1.7) or the equivalent version (1.9) to hold.

1.6 The common cause principle

The time asymmetry we found in the preceding section is not a completely satisfactory answer to the problem of characterizing the time asymmetry of probabilities. It is not clear that the macroscopic states such as $\mathcal{B}$ can be defined in such a way that the Markovian condition (1.7) is satisfied; moreover, there is the difficulty (pointed out to me by G. Sewell [22]) of proving consistency of the theory by showing that the dynamical consequences of the formula (1.7), studied in detail in [14] for isolated systems, are compatible with the dynamical laws. In the present section we look at a different approach, which uses microscopic histories rather than macroscopic events, and is much more easily reconciled with the dynamical laws.

In this discussion it will be assumed that there was an initial time, let us call it t_0, at which the Universe as we know it began. In order not to prejudge the time direction problem by making the intrinsic structure of the Universe temporally asymmetric quite apart from what is happening inside it, we suppose for the time being that there is will also be a final time t_1 at which the Universe as we know it will end. Such a time is indeed a feature of some cosmological models, see for example [21], although of course these models use curved space-time so that the straight light rays in our diagrams would have to be replaced by curved ones. We shall find that the value of t_1 plays no part in the discussion (except that it must be greater than t_0); so the results will apply equally well to a model universe in which $t_1 = +\infty$. The case $t_0 = -\infty$, which arises in the steady-state cosmological model, can also be treated (see [17]) but will be ignored here for simplicity.

The idea of referring back the present condition of the Universe, via deterministic mechanical laws, to its condition at time t_0 goes back at least to to Boltzmann who writes, when discussing the time asymmetry or "uni-directedness" of his H theorem, "The uni-directedness of this process is obviously not caused by the equations of motion of the molecules, for those do not change when the direction of time is changed. The uni-directedness lies uniquely and solely in the initial conditions. However, this is not to be understood in the sense that for each experiment one would have to make all over again the special assumption that the initial conditions are just particular ones and not the opposite, equally possible ones; rather, a unified fundamental assumption about the initial constitution of the world suffices, from which it follows automatically with logical necessity that whenever bodies engage in interaction then the correct initial conditions must prevail".[4]

Boltzmann's proposal for achieving this was to "conceive of the world as an enormously large mechanical system ... which starts from a completely ordered initial state, and even at present is still in a substantially ordered state" [5]. According to the usual interpretation of entropy as disorder, Boltzmann's remark means that the Universe started in a state of low entropy, and the entropy has been increasing ever since but is still quite low. Excellent elucidations of Boltzmann's proposal are given in refs [12,18]. Boltzmann's insight about the entropy of the initial state of the Universe is not the whole story, however. The law of increasing entropy is a property of certain processes in materials, and such proceses are only one item in our list of time-asymmetric processes in section 1.2. Moreover, it is hard to see how the value of just one number, the entropy, at the one time $t = t_0$ can control the subsequent evolution of the Universe with such exquisite precision as to determine the time direction of all the physical processes that happen everywhere for ever after. At the very least some information about the probabilities at time $t = t_0$ seems necessary.

As a step towards advancing Boltzmann's programme a bit further, I shall make use of Reichenbach's "principle of the common cause"[19] to formulate a reasonable hypothesis about the initial probabilities. Reichenbach's principle asserts that "if an improbable coincidence has occurred there must exist a common cause". Following Reichenbach himself, we may interpret "improbable coincidence" to mean simply a correlation, or more precisely a deviation from the product formula for the joint probability of two events or histories in two spatially separated space-time regions A and B. As for the "common cause", like all causes it takes place before its effect, and must therefore be some event in the past zones of both A and B, that is to say in their intersection $PZ(A) \cap PZ(B)$, which I shall call the *common past* of A and B. Thus, in the one-dimensional Universe illustrated in Fig. 5, if A and B are correlated, i.e. if $p_{A\cup B}(a \times b) \neq p_A(a)p_B(b)$, then Reichenbach's principle leads us to seek the cause of the correlation in some event that takes place in the region marked C.

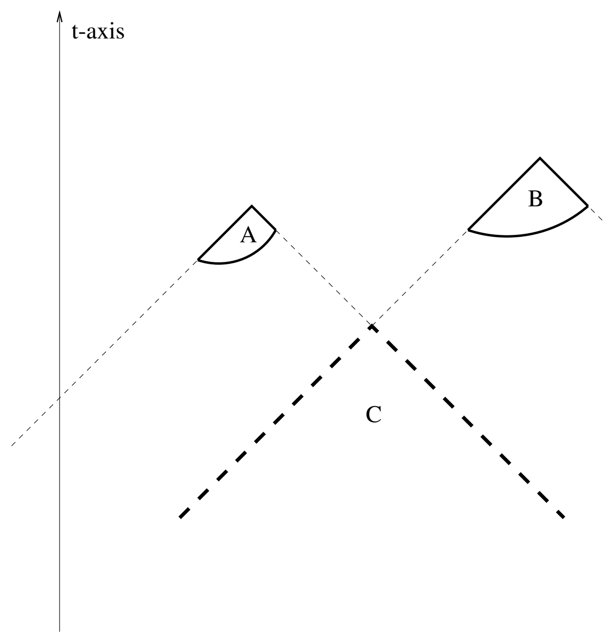

Fig. 5. Space-time regions for Reichenbach's common cause principle and the law of conditional independence. The region below the heavier dashed lines is $C = PZ(A) \cap PZ(B)$, the common past of A and B.

It is an elementary consequence of Reichenbach's principle that if the space-time regions A and B are so far apart that their past zones are disjoint, i.e. if $PZ(A) \cap PZ(B)$ is empty, as in Fig. 6, then the two regions are uncorrelated, that is to say the formula

$$p_{A\cup B}(a \times b) = p_A(a)p_B(b) \tag{1.10}$$

holds for all a in Ω_A and all b in Ω_B.

Eqn (1.10) can be applied to the case where the two "regions" are subsets of the manifold $t = t_0$, such as the two segments marked M and N in Fig. 6;

this leads us to the conclusion that disjoint parts of the $t = t_0$ manifold are uncorrelated[3].

From the independence of disjoint parts of the $t = t_0$ manifold we can derive a formula, which may be called[17] the *law of conditional independence*, relating the probabilities in two regions whose past zones are not disjoint. It states that if A and B are any two space-time regions, and C is the common past of A and B, then A and B are conditionally independent given the history of C. In symbols[4]

$$p_{A \cup B \cup C}(a \times b|c) = p_{A \cup C}(a|c) p_{B \cup C}(b|c) \tag{1.11}$$

for all $a \in \Omega_A, b \in \Omega_B, c \in \Omega_C$.

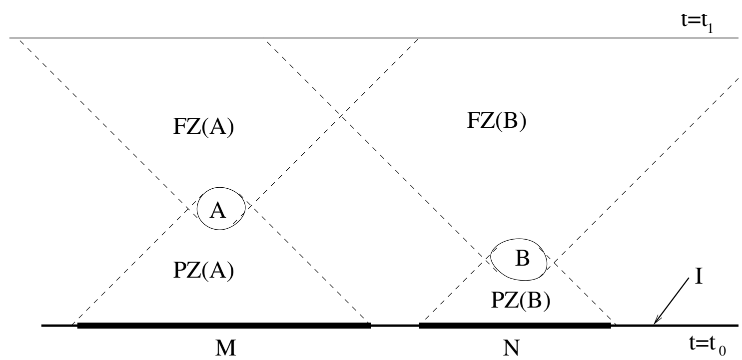

Fig. 6. The initial time is t_0, the final time t_1. The regions A and B are uncorrelated, because their past zones do not overlap (even though their future zones do overlap). For the same reason, the segments M and N are uncorrelated.

To prove (1.11), define I to be the initial manifold, on which $t = t_0$, and define two subsets M, N of I (see Fig. 7) by the formulas

$$\begin{aligned} M &= I \cap (PZ(A)) - C) \\ N &= I \cap (PZ(B)) - C), \end{aligned} \tag{1.12}$$

and let m, n denote their respective histories. By an argument similar to the one based on Fig. 6, the three regions M, N, C are uncorrelated with one another.

It can be seen from Fig. 7 that the definitions (1.12) etc. imply that $C \cup M$ severs the past zone of A, and that $C \cup N$ severs the past zone of B. Hence, by the determinism condition (1.1), there exist functions f, g such that

$$\begin{aligned} a &= f(m, c) \\ b &= g(n, c). \end{aligned} \tag{1.13}$$

[3] In [16] this independence of different parts of the $t = t_0$ manifold was taken as an axiom

[4] The same equation is given in [17] but the condition used there to characterize C appears to be too weak to ensure the truth of (1.11). Eqn (1.11), together with a diagram equivalent to Fig. 5, also appears in Bell's paper[1] about the impossibility of explaining quantum non-locality in terms of local "beables".

Applying the formula (1.3) we find that

$$\begin{aligned} p_{A\cup C}(a \times c) &= \sum_{m \in \Omega_M} p_{M\cup C}(m \times c)\delta(a, f(m,c)) \\ &= p_C(c) \sum_{m \in \Omega_M} p_M(m)\delta(a, f(m,c)) \end{aligned} \quad (1.14)$$

where in the last line we have used the fact that the regions M and C are uncorrelated. From (1.14) and the definition of conditional probability we obtain

$$p_{A\cup C}(a|c) = \sum_{m \in \Omega_M} p_M(m)\delta(a, f(m,c)). \quad (1.15)$$

Similar formulas can be worked out for $p_{B\cup C}(b|c)$ and $p_{A\cup B\cup C}(a \times b|c)$, and using all three formulas in (1.11) we find that the two sides of (1.11) are equal. This completes the proof.

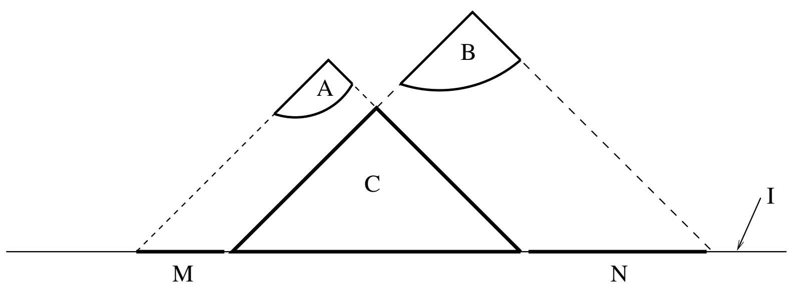

Fig. 7 Space-time regions used in proving the law of conditional independence.

Before going on to the extension of these ideas to quantum mechanics, a word about Bohmian mechanics[2] is in order. From our point of view, Bohmian mechanics is a deterministic classical theory, but since it contains simultaneous action at a distance the speed of light in our treatment would have to be taken infinite. This would make the dotted lines representing the light rays in the diagrams horizontal. Our formulation of Reichenbach's common cause principle would no longer work, but its consequence that disjoint pieces of the $t = t_0$ manifold are uncorrelated might still be adopted as an axiom in its own right. Initial conditions for Bohmian mechanics are treated in detail in [6], but with a different purpose in mind.

1.7 Quantum mechanics

Most of the main ideas in the model discussed here have analogues in quantum mechanics. Integration over a space of histories is an important tool in quantum-mechanical calculations, and so our history spaces Ω, Ω_A can be taken over directly into quantum mechanics. The probability distribution $p_A(a)$, however, is more problematical; its nearest analogue is a density matrix, a non-negative definite matrix which will be written $\varrho_A(a, a')$. In analogy with the dynamical condition on the classical probability distribution that dynamically impossible trajectories have zero probability, the density matrix also satisfies a dynamical condition. This condition is derivable from Schrödinger's equation, and has a property of symmetry under time reversal derivable from the fact that the complex conjugate of a wave function satisfies the time-reversed version of the Schrödinger equation. In relativistic quantum mechanics we would also expect it to have a property analogous to (1.3), enabling us to calculate the density matrix for a region A in terms of the one for a region B which severs the past zone of A. The analogue of (1.3) would have the form

$$\varrho_A(a, a') = \sum_{b \in \Omega_B} \sum_{b' \in \Omega_B} F_{AB}(a, b) F^\star_{AB}(a', b') \varrho_B(b, b') \tag{1.16}$$

where $F_{AB}(a, b)$ is a propagator and the star denotes a complex conjugate. The propagator may be given the following heuristic interpretation: if the history of the region B should happen to be b, then the wave function of A will be given by $\psi_A(a) = F_{AB}(a, b)$.

It is more difficult to find convincing quantum analogues for properties (ii) and (iii) of the classical probability model. The analogue of the shift invariance condition (1.4) is easy enough, simply $\varrho_{TA}(Ta, Ta') = \varrho_A(a, a')$. To give a precise formulation of the ergodic property (1.5), however, requires an understanding of what constitutes an "event" in quantum mechanics, a difficult problem[5] which is far beyond the scope of this article. Nevertheless we do know from experience that events do happen, and so the quantum mechanical model must be considered incomplete if it does not make some kind of provision for events.

Quantum mechanics contains a non-classical concept called *entanglement* which has no exact analogue in classical mechanics: it is formally similar to correlation, but more subtle. Two space-time regions A and B may be said to be entangled if their joint density matrix is not the product of the separate ones, i.e. if

$$\varrho_{A \cup B}(a \times b, a' \times b') \neq \varrho_A(a, b) \varrho_B(a', b'). \tag{1.17}$$

According to quantum theory, entanglement arises if a bipartite system is prepared in a suitable quantum state and the two parts then move apart from one another, as in the famous EPR paradox[7]. It therefore seems natural to adopt a quantum analogue of Reichenbach's principle, asserting that if two space-time regions are entangled there must exist a common cause. This common cause would be an event in the common past of A and B. If they have no common past, as in Fig. 6, then we would expect A and B to be unentangled.

If we want to formulate a quantum analogue of the law of conditional independence, we need some condition on the common past of the two regions A and B that will preclude the creation within it of any entanglement between A

[5] See, for example, [8,15].

and B. For example, it would presumably be sufficient to require C to be completely empty, both of matter and fields. More leniently, we could allow events to happen inside C provided that any incipient entanglement created thereby was destroyed before it got outside C. A possible way to achive this might be to make the requirement that whatever events take place in C they include something equivalent to making a complete measurement along the entire boundary of C. The measurement would intercept any particles or photons emerging from C and destroy any phase relations between their wave functions that might otherwise go on to generate entanglement between A and B.

1.8 Conclusion

Now we can look back at the various time-directed processes mentioned in section 1.2, to see whether their time direction can be derived from Reichenbach's common cause principle or whether it is logically independent from this principle and therefore in a sense accidental. In applying this principle, we shall normally identify the common cause with some interaction between the two correlated systems.

Recording devices, memory. A camera is essentially a closed box which interacts with the outside world for a short time while the shutter is open. The image on the film represents a correlation between the interior of the box and the world outside. According to the common cause principle, such a correlation implies an interaction in the common past; therefore the correlation (the image) can be there only after the shutter was opened, not before. The time direction of other recording devices, such as your memory, can be understood in the same way. For example, if you visit a new place, your memory of it is a correlation between your brain cells and the configuration of the place you went to. This correlation can exist only after the interaction between your body and the place during your visit; so you remember the visit after its occurrence, not before.

Irreversible processes in materials; increase of entropy. Boltzmann's original derivation of his kinetic equation for gases and the consequent H-theorem (the non-decrease of entropy in an isolated gas)[3] depended on his *Stosszahlansatz*, an assumption which says that the velocities of the gas molecules participating in a collision are uncorrelated prior to the collision. After the collision, on the other hand, they will in general be correlated. This is just the time direction we would expect from the the common cause principle applied to these collisions: the correlation comes after the interaction, not before.

In the rigorous derivation of Boltzmann's kinetic equation given by Lanford[9], the collisions are not treated individually but as part of the evolution of an isolated system of many interacting particles. In this case the time direction comes from Lanford's assumption that the particles are uncorrelated at the initial moment of the time evolution he studies (and, because of their subsequent interaction, only at that moment). The time direction given by this assumption can be seen to be consistent with the time direction of the common cause principle if we imagine the interaction to be switched on at this initial moment and then switched off again at some later moment; the common cause principle indicates that the particles are uncorrelated up until the moment when the interaction starts, but are (in general) correlated thereafter, even after the interaction has been switched off. Thus Lanford's assumption need not be seen as an *ad hoc* assumption for fixing the time direction in this particular problem, but instead as a particular case of the general scheme for fixing time directions provided by the common cause principle.

Other rigorous derivations of kinetic or hydrodynamic equations from reversible microscopic dynamics make similar assumption about an uncorrelated initial state, while not assuming anything in particular about the final state, and so their time direction can be understood on the basis the common cause principle in the same way as for Lanford's result. For example, Lebowitz and Spohn[10,11], deriving Fick's law for self-diffusion in a gas consisting of hard spheres of two different colours, assume that the the colours of the particles are initially uncorrelated with the dynamical states of the particles.

Boltzmann also gave a more general argument (for an enthusiastic explanation see [12,13]) for the increase of entropy in an isolated system, which does not use any detailed assumptions about the nature of the interactions, nor does it assume that the particles are uncorrelated at the moment when the evolution process under study begins, which is normally a moment when the system becomes isolated. Instead he assumes that the entropy of the system at this initial time is less than the equilibrium value of the entropy, and one can then argue plausibly that the entropy is likely to move closer to the equilbrium value. i.e. to increase. But what grounds do we have to assume that the lower value of entropy occurs at the moment when the system becomes isolated rather than the later moment when the system ceases to be isolated? Boltzmann's suggestion to base everything on the assumption of a very low initial entropy for the Universe at its initial time $t = t_0$ is very important, but to me it is not a complete answer to the question, since there is no obvious reason why the entropy of a small temporarily isolated part of the universe has to vary with time in the same direction as the entropy of the universe as a whole.

Once again, the common cause principle suggests an answer. The problem is to understand why an isolated system is closer to equilibrium at the end of its period of isolation than at the beginning. When an isolated system is in equilibrium, its correlation with its surroundings is the least possible compatible with its values for the thermodynamic parameters such as energy; whereas if it is not in equilibrium it is more strongly correlated with its surroundings[6]. By the common cause principle, such correlations imply a past interaction between

[6] Such correlations can be regarded as "information" held by the surroundings, which may include human observers, about the system. Indeed there is a quantitative relation, due to L. Szilard, between the amount of this information and the entropy of the system. (For a detailed treatment of this relation, with some references to earlier work, see pp 226-231 of [14].) As time progresses, this information loses its relevance to the current state of the isolated system, the amount of correlation goes down, and the entropy of the system goes up.

system and surroundings; therefore we would expect the correlations to exist at the beginning of the period of isolation when the system has just been interacting with its surroundings, rather than at the end when it has not. (For example, the system might have been prepared in some non-equilibrium state by an experimenter: the correlation between system and experimenter implied by the non-equilibrium state – and the experimenter's knowledge of it – arose from a prior interaction.) In this way the common cause principle provides a rational explanation of why the low-entropy state, in which the system is more strongly correlated with is surroundings, occurs at the beginning of the period of isolation rather than at the end.

Expanding waves. Different points on a spherical wave are correlated. By the common cause principle, this correlation was caused by a previous interaction, in this case a local interaction with an electron in the antenna; therefore the wave comes after the local event that produces it, and must expand rather than contract.

Expanding universe. This is the one item in our list whose time direction clearly does not follow from the common cause principle. It is true that the expansion we see, a correlation between the motions of distant galaxies, implies a past interaction, presumably that which took place at the time of the Big Bang. But a contracting Universe would also be compatible with the common cause principle, since the correlations of the galactic motions which constituted the contraction could still be attributed to a past interaction, namely the long-range gravitational attraction between the galaxies.

Black and white holes. A proper treatment of this subject is beyond the scope of this paper; to produce one we would need to regard the metric of space-time as part of the kinematic description of the Universe instead of regarding it as a fixed background within which the rest of the kinematics takes place. All we can do here is to indicate a few hints that can be obtained by treating black or white holes as part of the background rather than part of the dynamics.

A black hole arises when a very heavy star is no longer hot enough to support itself against its self-gravitation. It is a singularity in curved space-time which begins at the time of collapse and (in classical gravitation theory) remains for ever thereafter. It is called a "black" hole because (again in classical theory) no light, or anything else, comes out of it. Given any point in the non-singular part of space-time, whether inside or outside the event horizon, all the light reaching that point comes from places other than the black hole. The past zones of space-time points near a black hole are bent by the strong graviational field, but topologically they are essentially the same as the ones in Figs. 5 and 7, and so there is no inconsistency with the common cause principle.

The exact time inverse of a black hole would be a white hole that started at the beginning of time and disappeared at a certain moment. Given a point in space-time sufficiently close to the white hole, some or all of the light and other causal influences reaching it come from the white hole rather than from the $t = t_0$ manifold. The past zones of points near the white hole therefore need not extend back to the $t = t_0$ manifold, but may instead end on the white hole itself. The common cause principle applied to this situation would lead us to conclusions about the white hole similar to the ones reached in section 1.6 about the initial manifold, different pieces of its "surface" (to the extent that a line singularity in space-time can be said to have a surface) being uncorrelated, both with each other and with the $t = t_0$ manifold. It seems, then, that the existence of white holes would be consistent with the common cause principle. The enormous gravitational forces at the surface of a white hole would no doubt have a profound effect on whatever came out of it; indeed one could speculate that the surface of a white hole is at an infinite temperature, or even hotter than that[7].

Acknowledgements

I am much indebted to Ian Percival for many discussions and ideas. I am also indebted to the Societa Italiana di Fondamenti della Fisica, the Istituto Italiano per gli Studi Filosofici and the Royal Society of London for Improving Natural Knowledge for financial support enabling me to attend the conference and talk about this work. I am grateful to Michael Kiessling for translating a difficult passage from Boltzmann's book into English.

References

1. J. S. Bell: 'The theory of local beables', *Speakable and unspeakable in quantum mechanics* (Cambridge University Press 1987), pp 52-62
2. D. Bohm: 'A Suggested Interpretation of the Quantum Theory in terms of "Hidden Variables"', Phys. Rev. **85** 166-179 and 180-193 (1952)
3. L. Boltzmann: *Vorlesungen über Gastheorie* (J. A. Barth, Leipzig, 1896 and 1898), section 5. English translation by S. G. Brush, *Lectures on gas theory* (University of California Press, Berkeley and Los Angeles, 1964)
4. L. Boltzmann: ibid., section 87. English translation by Michael Kiessling (private communication, 2000)
5. L. Boltzmann: ibid., section 89. English translation by S. G. Brush, loc. cit.
6. D. Dürr, S. Goldstein and N. Zanghi: 'Quantum Equilibrium and the Origin of Absolute Uncertainty', J. Stat. Phys. **67** 843-907 (1992)
7. A. Einstein, P. Podolsky and N. Rosen: 'Can Quantum-mechanical Description of Reality be Considered Complete?', Phys. Rev. **47**, 777-780 (1935)
8. R. Haag: 'Fundamental irreversibility and the concept of events', Commun. Math. Phys.**132**, 245-251 (1990)
9. O. Lanford III: 'Time Evolution of Large Classical Systems'. In: *Dynamical Systems: Theory and Applications: Battelle Seattle Rencontres 1974*, ed. by J. Moser (Springer Lecture Notes in Physics **38**, 1975) pp.1-111
10. J. L. Lebowitz and H. Spohn: 'Microscopic Basis for Fick's Law of Self-difusion', J. Stat. Phys. **28** 539-556 (1982)

[7] The so-called "negative temperatures" are hotter than any positive or even infinite temperature, in the sense that energy flows from any negative temperature to any positive one.

11. J. L. Lebowitz and H. Spohn: ‘On the Time Evolution of Macroscopic Systems’, Commun. Pure Appl. Math. **36** 595-613 (1983)
12. J. L. Lebowitz: ‘Boltzmann’s Entropy and Time’s Arrow’, Physics today **46:9** 32-38 (1993)
13. J. L. Lebowitz: ‘Statistical Mechanics: A Selective Review of Two Central Issues’, Rev. Mod. Phys. **71**, S346-357 (1999)
14. O. Penrose: *Foundations of statistical mechanics* (Pergamon, Oxford, 1970)
15. O. Penrose: ‘Quantum Mechanics and Real Events’. In: *Quantum Chaos – Quantum Measurements*, ed. P. Cvitanović et al. (Kluwer, Netherlands 1992) pp 257-264
16. O. Penrose: ‘The “Game of Everything”’. Markov Processes and Related Fields **2**, 167-182 (1996)
17. O. Penrose and I. C. Percival: ‘The Direction of Time’, Proc. Phys. Soc. **79**, 605-616 (1962)
18. R. Penrose: *The Emperor’s New Mind* (Oxford University Press, Oxford 1989), chapter 7
19. H. Reichenbach: *The Direction of Time* (University of California Press, Berkeley 1956), especially section 19 .
20. L. S. Schulman: Opposite Thermodynamic Arrows of Time, Phys. Rev. Letters **83** 5419-5422 (1999)
21. D. W. Sciama: *Modern Cosmology and the Dark Matter Problem* (Cambridge University Press, Cambridge 1993), section 3.2
22. G. Sewell: private communication (1970)