The logic of the past hypothesis

David Wallace

July 15, 2011

Abstract

I attempt to get as clear as possible on the chain of reasoning by which irreversible macrodynamics is derivable from time-reversible microphysics, and in particular to clarify just what kinds of assumptions about the initial state of the universe, and about the nature of the microdynamics, are needed in these derivations. I conclude that while a "Past Hypothesis" about the early Universe does seem necessary to carry out such derivations, that Hypothesis is not correctly understood as a constraint on the early Universe's entropy.

1 Introduction

There are no consensus positions in philosophy of statistical mechanics, but the position that David Albert eloquently defends in *Time and Chance* (Albert 2000) is about as close as we can get.[1] It hinges on two views (in *Time and Chance*, the latter gets most of the air time, but both play crucial roles):

1. The tendency of systems' entropy to increase is basically just a consequence of the geometry of phase space. That region of phase space corresponding to a system being at equilibrium is so very large compared to the rest of phase space that unless either the dynamics or the initial state are (as Goldstein (2001) puts it) "ridiculously special", then the system will in fairly short order end up in the equilibrium region.

2. The observed asymmetry in statistical mechanics — in particular, the tendency of entropy to increase rather than decrease — can be derived from time-symmetric microphysics provided we are willing to postulate that the entropy of the early universe is very low compared to the current entropy of the universe — what Albert has memorably dubbed the "Past Hypothesis".

There is something rather puzzling about both views. Take the first: it seems to suggest that any given system, unless it is "ridiculously special", will quickly

[1] Albert-like claims are also espoused by, e. g. , Goldstein (2001), Lebowitz (2007), Callender (2009), and Penrose (1989, 2004).

end up in equilibrium. But of course, in the real world, we very frequently find systems far from equilibrium — indeed, life itself depends on it. And many of those systems, even when isolated from their surroundings, refuse to evolve into equilibrium. A room filled with a mixture of hydrogen and oxygen, at room temperature, can remain in that state for years or decades, yet one has only to strike a spark in that room to be reminded that it is not an equilibrium state. Indeed, a room filled with hydrogen at room temperature is not really at equilibrium: it is thermodynamically favourable for it to fuse into iron, but you would wait a long time for this to happen.

Furthermore, we have a detailed, quantitative understanding of exactly how quickly systems in various non-equilibrium states evolve towards equilibrium. In particular, chemists (whether of the ordinary or nuclear variety) have precise and thoroughly tested dynamical theories which predict, from the microdynamics, just how quickly systems complete their irreversible movement towards equilibrium. It is, at best, very difficult to see how these quantitative theories of the approach to equilibrium fit into the very general argument for equilibration given by Albert, Goldstein *et al.*

The Past Hypothesis is puzzling in a different way. It suggests, or seems to suggest, that our knowledge of the low entropy of the early universe is somehow special: we are not supposed to know the Past Hypothesis in the way we usually know information about the past, but rather, we are justified in postulating it because without that postulate, all of our beliefs about the past would be unjustified. But there is something a little odd here: after all, we have (or think we have) rather detailed knowledge of the macroscopic state of the early Universe gained from cosmology, and we can calculate its entropy fairly accurately. It is also not clear (see in particular the trenchant criticisms of Earman (2006)) exactly how imposing a low entropy at the beginning of time can lead to irreversible physics here and now.

And yet... for all that, there is clearly something to both views. There does seem to be some important sense in which irreversibility is connected with phase space volume and the behaviour of typical — that is, not-ridiculously-special — systems. And it does seem that, absent time-asymmetry in microphysics, as a matter of logic there must be some link between the boundary conditions of the Universe and the observed time-asymmetry in macrophysics.

My purpose in this paper is to try to get as clear as possible on just how the logic of deriving macrophysical irreversibility from microdynamics-plus-past-hypothesis is supposed to go. My starting point is the observation above: that we actually have a large body of quantitative theory about irreversible physical processes, and any adequate account of irreversibility needs to explain the quantitative success of these theories and not just the qualitative observation that systems tend to equilibrium. So in sections 2–4 I set aside philosophical concerns and try to get as clear as possible on what the mathematical route is by which we derive empirically reliable irreversible macrodynamics from reversible microdynamics. In sections 5–6 I examine just when this mathematical route is physically justified, and conclude that a Past Hypothesis is indeed needed, but of a rather different character from what is usually argued. I conclude by making contact again with the two views mentioned above, and in particular with Albert's own approach.

I should draw attention to two distinctive features of my approach. Firstly, one frequent theme in criticism of the two views given above has been on their lack of mathematical rigor and care (see, in particular, Frigg's criticism of Goldstein (Frigg 2008) and Earman's objections (Earman 2006) to any assignment of entropy to the early Universe). By contrast, I am perfectly happy to allow their proponents to make whatever plausible-sounding mathematical conjectures they like (and indeed, make several such myself in my own account). My concern, rather, is in understanding just what those conjectures are supposed to achieve and why they can be expected to achieve it. The purpose of the philosopher of physics, it might be argued, is not to prove theorems but to see which theorems are worth proving.

Secondly, it seems all but universal to conduct discussions of statistical mechanics at the classical level. Sklar's account of the reasons for this appears to be fairly typical:

> [T]he particular conceptual problems on which we focus — the origin and rationale of probability distributions over initial states, the justification of irreversible kinetic equations on the basis of reversible underlying dynamical equations, and so on — appear, for the most part, in similar guise in the development of both the classical and quantum versions of the theory. The hope is that by exploring these issues in the technically simpler classical case, insights will be gained that will carry over to the understanding of the corrected version of the theory. ...This way of doing things is not idiosyncratic, but common in the physics literature devoted to foundational issues. (Sklar 1993, p.12).

But I am not convinced that the classical case really is "technically simpler" (at least where study of general features of the theory, rather than rigorous analysis of specific systems, is our goal), and nor am I confident that the conceptual problems really do appear "in similar guise". Notably, quantum mechanics contains probability at an essential level; it also includes its own form of irreversibility in the form of decoherence-induced branching. So my approach is in general to study the classical and quantum cases in parallel, and to neglect the classical theory in favour of the quantum one where they differ in important respects. If we are interested in understanding irreversibility in our world, after all, classical systems should be of interest to us only insofar as they are good approximations to quantum systems.

In discussing quantum mechanics, I assume that (i) the quantum state is physically real, (ii) it evolves unitarily at all times, and (iii) there are no hidden variables. That is, I basically assume the Everett interpretation (discussed and developed *in extenso* in Wallace (2011) and Saunders, Barrett, Kent, and Wallace (2010)). In doing so, of course, I part company with Albert: *Time and Chance* is an admirable exception to the usual classical-physics-only trend, but its quantum-mechanical discussions are largely confined to explicitly time-asymmetric dynamical-collapse theories. Much of what I say should, however, carry over to versions of quantum theory with hidden variables of one kind or another, such as modal interpretations or pilot-wave theories.

2 The macropredictions of microdynamics

For present purposes, classical and quantum mechanics have, essentially, a similar dynamical form. In both cases, we have

- A state space (phase space or (projective) Hilbert space);
- A deterministic rule determining how a point on that state space evolves over time (generated by the classical Hamiltonian and the symplectic structure, or the quantum Hamiltonian and the Hilbert-space structure, as appropriate);
- Time reversibility, in the sense that given the state at time t, the dynamics is just as well suited to determine the state for times before t as for times after t.

I will also assume that, whatever particular version of each theory we are working with, both theories have something which can reasonably be called a "time reversal" operator. This is a map τ from the state space to itself, such that if the t-second evolution of x is y then the t-second evolution of τy is τx; or, equivalently, if $x(t)$ solves the dynamical equations then so does $\tau x(-t)$. I'm not going to attempt a formal criterion for when something counts as a time-reversal operator; in classical and quantum mechanics, we know it when we see it. (Though in quantum field theory, it is the transformation called CPT, and not the one usually called T, that deserves the name).

Both theories also have what might be called, neutrally, an "ensemble" or "distributional" variant, though here they differ somewhat. In the classical case, the deterministic dynamics induces a deterministic rule to evolve functions over phase space, and not just points on phase space: if the dynamical law is given schematically by a function φ_t, so that $\varphi_t(x)$ is the t-second evolution of x, then $\varphi_{t*}\rho = \rho \cdot \varphi_t$. In more concrete and familiar terms, this takes us from the Hamiltonian equations of motion for individual systems to the Liouvillian equations for ensembles.

In the quantum case, we instead transfer the dynamics from pure to mixed states. If the t-second evolution takes state $|\psi\rangle$ to $\widehat{U}_t\,|\psi\rangle$, the distributional variant takes ρ to $\widehat{U}_t\rho\widehat{U}^\dagger_t$.

I stress: the existence of these distributional variants is a purely mathematical claim; no statement of their physical status has yet been made. The space of functions on, or density operators over, the state space can be thought of, mathematically speaking, as a state space in its own right, for the distributional variant of the theory.

In principle, the way we use these theories to make predictions ought to be simple: if we want to know the state of the system we're studying in t seconds' time, we just start with its state now and evolve it forward for t seconds under the microdynamics. And similarly, if we want to know its state t seconds ago, we just time-reverse it, evolve it forward for t seconds, and time-reverse it again. (Or equivalently, we just evolve it forwards for $-t$ seconds.)

And sometimes, that's what we do in practice too. When we use classical mechanics to predict the trajectory of a cannonball or the orbit of a planet, or when we apply quantum mechanics to some highly controlled situation (say, a quantum computer), we really are just evolving a known state under a known dynamics. But of course, in the great majority of situations this is not the case, and we have to apply approximation methods. Sometimes that's glossed as being because of our lack of knowledge of the initial state, or our inability to solve the dynamical equations exactly, but this is really only half the story. Even if we were able to calculate (say) the expansion of a gas in terms of the motions of all its myriad constituents, we would have missed important generalisations about the gas by peering too myopically at its microscopic state. We would, that is, have missed important, robust higher-level generalisations about the gas. And in quantum mechanics, the emergent behaviour is frequently the only one that physically realistic observers can have epistemic access to: decoherence strongly constrains our ability to see genuinely unitary dynamical processes, because it's too difficult to avoid getting entangled with those same processes.

The point is that in general we are not interested in all the microscopic details of the systems we study, but only in the behaviour of certain more coarse-grained details. It is possible (if, perhaps, slightly idealised) to give a rather general language in which to talk about this: suppose that $t_1, \ldots t_N$ is an increasing sequence of times, then a set of *macroproperties* for that sequence is an allocation, to each time t_i in the sequence, of either

(i) in the classical case, a Boolean algebra of subsets of the system's phase space whose union is the entire phase space; or

(ii) in the quantum case, a Boolean algebra of subspaces of the system's Hilbert space whose direct sum is the entire Hilbert space.

In both cases, it is normal to specify the macroproperties as being unions or direct sums (as appropriate) of macro*states*: a set of macrostates for a (classical / quantum) system is a set of mutually (disjoint / orthogonal) (subsets / subspaces) whose (union / direct sum) is the entire state space. Throughout this paper, I will assume that any given set of macroproperties is indeed generated from some set of macrostates in this way. (And in most *practical* cases, the choice of macrostate is time-independent.) For the sake of a unified notation, I will use $\oplus$ to denote the union operation for classical sets and the direct sum operation for quantum subspaces, $\subset$ to denote the subset relation for classical sets and the subspace relation for quantum subspaces, and "disjoint" to mean either set-theoretic disjointness or orthogonality, as appropriate.

The idea of this formalism is that knowing that a system has a given macroproperty at time t_i gives us some information about the system's properties at that time, but only of a somewhat coarse-grained kind. We define a *macrohistory* α of a system as a specification, at each time t_i, of a macroproperty $\alpha(t_i)$ for that time; the set of all macrohistories for a given set of macroproperties is the *macrohistory space* for that set.. It should be fairly clear that given the macrohistory space of a given set of macroproperties, we can recover that set; hence I speak interchangably of a macrohistory space for a theory and a set of macroproperties for the same theory. For simplicity, I usually drop the 'macro' qualifier where this is not likely to cause confusion.

A few definitions: by a history of length K (where $K < N$) I mean a history which assigns the whole state space to all times t_i for $i > M$. Given histories α and β of lengths K and K' (with $K < K'$) then α is an *initial segment* of β if $\alpha(t_i) = \beta(t_i)$ for $i \leq M$. Given macrohistories α and β, we can say that α is a *coarsening* of β if $\beta(t_i) \subset \alpha(t_i)$ for each time t_i at which they are defined, and that α and β are *disjoint* if $\beta(t_i)$ and $\alpha(t_i)$ are disjoint at each t_i. A history β is the *sum* of a (countable) set of mutually disjoint histories $\{\alpha_j\}$ (write $\beta = \oplus_j \alpha_j$)if $\beta(t_i) = \oplus_j \alpha_j(t_i)$ for all t_i, and, in particular, a set of disjoint histories is *complete* if their sum is the trivial history $\widehat{1}$ whose macroproperty at each time is just the whole state space. And a *probability measure* Pr for a given history space is a real function from histories to $[0, 1]$ such that

1. If $\{\alpha_j\}$ is a countable set of disjoint histories then $\Pr(\oplus_j \alpha_j) = \sum_j \Pr(\alpha_j)$, and

2. $\Pr(\widehat{1}) = 1$.

The point of a probability measure over a history space is that it determines a (generally stochastic) dynamics: given two histories α and β where α is an initial segment of β, we can define the *transition probability* from α to β as $\Pr(\beta)/\Pr(\alpha)$. A *macrodynamics* for a (classical or quantum) system is then just a history space for that system, combined with a probability measure over that history space. A macrodynamics is *branching* iff whenever α and β agree after some time t_m but disagree at some earlier time, either $\Pr(\alpha) = 0$ or $\Pr(\beta) = 0$; it is *deterministic* if whenever α and β agree before some time t_m but disagree at some later time, either $\Pr(\alpha) = 0$ or $\Pr(\beta) = 0$.

With this formalism in place, we can consider how classical and quantum physics can actually induce macrodynamics: that is, when it will be true, given the known microdynamics, that the system's macroproperties obey a given macrodynamics. The simplest case is classical mechanics in its non-distributional form: any given point x in phase space will have a determinate macrostate at any given time, and so induces a deterministic macrodynamics: if $U(t) \cdot x$ is the $t-$second evolution of x under the classical microdynamics, then

$$\begin{aligned} \Pr_x(\alpha) &= 1 \quad (\text{if } U(t_n - t_1) \cdot x \in \alpha(t_n) \text{ for all } n) \\ \Pr_x(\alpha) &= 0 \qquad\qquad\quad (\text{otherwise}) \end{aligned} \tag{1}$$

To get stochastic dynamics from classical microdynamics, we need to consider the distributional version. Suppose that at time t_1 the probability of the system having state x is $\rho(x)$; then the probability at time t_n of it having state x is given by evolving ρ forward for a time $t_n - t_1$ under the distributional (Liouville) dynamics. Writing $L(t) \cdot \rho$ for the $t-$second evolution of ρ and $P(M) \cdot \rho$ for the restriction of ρ to the macrostate M, we define the *history super-operator* $H(\alpha)$ by

$$H(\alpha)\cdot\rho = P(\alpha(t_n))\cdot L(t_n - t_{n-1})\cdot P(\alpha(t_{n-1}))L(t_{n-1} - t_{n-2} \cdots L(t_2 - t_1)P(\alpha(t_1))\cdot\rho. \tag{2}$$

$H(\alpha) \cdot \rho$ is the distribution obtained by alternately evolving ρ forward and then restricting to the successive terms in α. So we have that the probability of history α given initial distribution ρ is

$$\Pr_\rho(\alpha) = \int H(\alpha) \cdot \rho \tag{3}$$

where the integral is over all of phase space.s

A formally similar expression can be written in quantum mechanics. There, we write ρ for the system's density operator at time t_1, $L(t) \cdot \rho$ for the t-second evolution of ρ under the unitary dynamics (so if $\widehat{U}(t)$ is the t-second unitary time translation operator, $L(t) \cdot \rho = \widehat{U}(t)\rho\widehat{U}^\dagger(t))$, and $P(M) \cdot \rho$ for the projection of ρ onto the subspace M (so that if $\widehat{\Pi}_M$ is the standard projection onto that subspace, $P(M) \cdot \rho = \widehat{\Pi}_M \rho \widehat{\Pi}_M$). Then (2) can be understood quantum-mechanically, and (3) becomes

$$\Pr_\rho(\alpha) = \mathsf{Tr}(H(\alpha) \cdot \rho). \tag{4}$$

The resemblance is somewhat misleading, however. For one thing, in classical physics the macrodynamics are probabilistic because we put the probabilities in by hand, in the initial distribution ρ. But in quantum physics, (4) generates stochastic dynamics even for the pure-state version of quantum theory (relying on Part II to explain why the weights of histories deserve to be called "probabilities"). And for another, (4) only defines a probability measure in special circumstances. For if we define the *history operator* $C(\alpha)$ by

$$\widehat{C}(\alpha) = \Pi_{\alpha_n}\widehat{U}(t_n - t_{n-1})\Pi_{\alpha_{n-1}} \cdots \widehat{U}(t_2 - t_1)\Pi(\alpha_1), \tag{5}$$

we can express $H(\alpha)$ by

$$H(\alpha) \cdot \rho = \widehat{C}(\alpha)\rho\widehat{C}^\dagger(\alpha) \tag{6}$$

and rewrite (4) as

$$\Pr_\rho(\alpha) = \mathsf{Tr}(\widehat{C}(\alpha)\rho\widehat{C}^\dagger(\alpha)), \tag{7}$$

in which case

$$\Pr_\rho(\sum_j \alpha_j) = \sum_{j,k} \mathsf{Tr}(\widehat{C}(\alpha_j)\rho\widehat{C}^\dagger(\alpha_k)), \tag{8}$$

which in general violates the requirement that $\Pr_\rho(\sum_j \alpha_j) = \sum_j \Pr_\rho(\alpha_j)$. To ensure that this requirement is satisfied, we need to require that the history space satisfies the decoherence condition: that the decoherence function

$$d_\rho(\alpha, \beta) \equiv \mathsf{Tr}(\widehat{C}(\alpha)\rho\widehat{C}^\dagger(\beta)) \tag{9}$$

vanishes unless α is a coarsening of β. (A weaker requirement — that the real part of the decoherence functional vanishes — would be formally sufficient but seems to lack physical significance.) In general, this is ensured in practical examples by environment-induced decoherence (cf Wallace (2011, chapter 3) and references therein for further discussion).

Before moving on, I should stress that the entire concept of a history operator, as defined here, builds in a notion of time-asymmetry: by construction, we have used the system's distribution at the initial time t_1 to generate a probability measure over histories defined at that and all subsequent times. However, we could equally well have defined histories running backwards in time — 'antihistories', if you like — and used the same formalism to define probabilities over antihistories given a distribution at the final time for those antihistories.

3 Coarse-grained dynamics

The discussion so far has dealt entirely with how macroscopic dynamics can be extracted from the microscopic equations, assuming that the latter have been solved exactly. That is, the framework is essentially descriptive: it provides no shortcut to determining what the macrodynamics actually are. In reality, though, it is almost never the case that we have access to the exact micro-level solutions to a theory's dynamical equations; instead, we resort to certain approximation schemes both to make general claims about systems' macrodynamics and to produce closed-form equations for the macrodynamics of specific systems. In this section, I wish to set out what I believe to be *mathematically* going on in these approximation schemes, and what assumptions of a purely technical nature need to be made. For now, I set aside philosophical and conceptual questions, and ask the reader to do likewise.

The procedure we use is intended to allow for the fact that we are often significantly ignorant of, and significantly uninterested in, the microscopic details of the system, and instead wish to gain information of a more coarse-grained nature, and it seems to go like this. Firstly, we identify a set of macroproperties (defined as above) in whose evolution we are interested. Secondly, we define a map $\mathcal{C}$ — the *coarse-graining map* — which projects from the distribution space onto some subset S_C of the distributions. By "projection" I mean that $\mathcal{C}^2 = \mathcal{C}$, so that the distributions in S_C — the "coarse-grained" distributions — are unchanged by the map. It is essential to the idea of this map that it leaves the macroproperties (approximately) unchanged — or, more precisely, that the probability of any given macroproperty being possessed by the system is approximately unchanged by the coarse-graining map. In mathematical terms, this translates to the requirement that for any macroproperty M,

$$\int_M \mathcal{C}(\rho) = \int_M \rho \tag{10}$$

in the classical case, and

$$\mathsf{Tr}(\Pi_M \mathcal{C}(\rho)) = \mathsf{Tr}(\Pi_M \rho) \tag{11}$$

in the quantum case. I will also require that $\mathcal{C}$ commutes with the time reversal operation (so that the coarse-graining of a time-reversed distribution is the time-reverse of the coarse-graining of the distribution).

We then define the *forward dynamics induced by* $\mathcal{C}$ — or the $\mathcal{C}+$ dynamics for short — as follows: take any distribution, coarse-grain it, time-evolve it forward (using the microdynamics) by some small time interval Δt, coarse-grain it again, time-evolve it for another Δt, and so on. (Strictly speaking, then, Δt ought to included in the specification of the forward dynamics. However, in practice, we are only interested in systems where (within some appropriate range) the induced dynamics are insensitive to the exact value of Δ_t.)

By a *forward dynamical trajectory induced by* $\mathcal{C}$, I mean a map from (t_i, ∞) into the coarse-grained distributions (for some t_i), such that the distribution at t_2 is obtained from the distribution at t_1 by applying the $\mathcal{C}+$ dynamics whenever $t_2 > t_1$. A *section* of this trajectory is just a restriction of this map to some finite interval $[t, t']$.

What is the coarse-graining map? It varies from case to case, but some of the most common examples are

The coarse-grained exemplar rule: Construct equivalence classes of distributions: two distributions are equivalent if they generate the same probability function over macroproperties. Pick one element in each equivalence class, and let the coarse-graining map take all elements of the equivalence class onto that element. This defines a coarse-graining rule in classical or quantum physics; in practice, however, although it is often used in foundational discussions, rather few actual applications make use of it.

The measurement rule: Replace the distribution with the distribution obtained by a nonselective measurement of the macrostate: that is, apply

$$\rho \to \sum_M \Pi_M \rho \Pi_M \tag{12}$$

where the sum ranges over macrostates.[2] (This obviously only counts as a coarse-graining in quantum mechanics; the analogous classical version, where ρ is replaced by the sum of its restrictions to the macrostates, would be trivial.)

[2] To avoid problems with the quantum Zeno effect (Misra and Sudarshan (1977); see Home and Whitaker (1997) for a review) for very small δt, the measurement rule strictly speaking ought to be slightly unsharpened (for instance, by using some POVM formalism rather than sharp projections onto subspaces); the details of this do not matter for our purposes.

The correlation-discard rule: Decompose the system's state space into either the Cartesian product (in classical physics) or the tensor product (in quantum physics) of state spaces of subsystems. Replace the distribution with that distribution obtained by discarding the correlations between subsystems (by replacing the distribution with the product of its marginals or the tensor product of its partial traces, as appropriate).

One note of caution: the correlation-discard rule, though very commonly used in physics, will fail to properly define a coarse-graining map if the probability distribution over macroproperties itself contains nontrivial correlations between subsystems. In practice this only leads to problems if the system does not behave deterministically at the macroscopic level, so that such correlations can develop from initially uncorrelated starting states. Where this occurs, the correlation-discard rule needs generalising: decompose the distribution into its projections onto macrostates, discard correlations of these macrostates individually, and re-sum. Note, though, that in quantum mechanics this means that two coarse-grainings are being applied: to "decompose the distribution into its projections onto macrostates and then re-sum" is just to perform a non-selective measurement on it — that is, to apply the measurement rule for coarse-grainings.

Another example is again often used in foundational discussions of statistical mechanics, but turns up rather less often in practical applications:

The smearing rule: Blur the fine structure of the distribution by the map

$$\rho' = \int \mathrm{d}q'\,\mathrm{d}p'\,f(q',p')T(q',p')\cdot\rho \tag{13}$$

where $T(q',p')$ is translation by $(q'p')$ in phase space and f is some function satisfying $\int f = 1$ and whose macroscopic spread is small. A simple choice, for instance, would be to take f to be a suitably-normalised Gaussian function, so that

$$\rho' = \mathcal{N}\int \mathrm{d}q'\,\mathrm{d}p'\,\exp[-(q-q')^2/(\Delta q^2)]\exp[-(p-p')^2/(\Delta p^2)]\rho(q,p) \tag{14}$$

where ρ is to be read as either the phase-space probability distribution (classical case) or the Wigner-function representation of the density operator (quantum case).

For a given system of $\mathcal{C}+$ dynamics, I will call a distribution *stationary* if its forward time evolution, for all times, is itself. (So stationary distributions are always coarse-grained.) Classic examples of stationary distributions are the (classical or quantum) canonical and microcanonical ensembles. Distributions involving energy flow (such as those used to describe stars) look stationary, but generally aren't, as the energy eventually runs out.

How do we generate empirical predictions from the coarse-grained dynamics? In many cases this is straightforward, because those dynamics are deterministic at the macroscopic level ("macrodeterministic"): if we begin with a coarse-grained distribution localised in one macrostate, the $\mathcal{C}+$ dynamics carries it into a coarse-grained distribution still localised in one (possibly different) macrostate.

More generally, though, what we want to know is: how probable is any given sequence of macrostates? That is, we need to apply the history framework used in the previous section. All this requires is for us to replace the (in-practice-impossible-to-calculate) macrodynamics induced by the microdynamics with the coarse-grained dynamics: if $L^{C+}(t)\cdot\rho$ is the $t-$second evolution of ρ under the $\mathcal{C}+$-dynamics, and $P(M)\cdot\rho$ is again projection of ρ onto the macroproperty M, then we can construct the coarse-grained history superoperator

$$H^{C+}(\alpha)=P(\alpha(t_n))\cdot L^{C+}(t_n-t_{n-1})\cdot P(\alpha(t_{n-1}))\cdot L^{C+}(t_{n-1}-t_{n-2})\cdots L^{C+}(t_2-t_1)\cdot P(\alpha(t_1)). \tag{15}$$

(It should be pointed out for clarity that each $L^{C+}(t_k-t_{k-1})$ typically involves the successive application of many coarse-graining operators, alternating with evolution under the fine-grained dynamics; put another way, typically $t_k-t_{k-1}\gg\Delta t$. Even for the process to be well-defined, we have to have $t_k-t_{k-1}\geq\Delta t$; in the limiting case where $t_k-t_{k-1}=\Delta t$, we obtain $H^{C+}(\alpha)$ by alternately applying *three* operations: evolve, coarse-grain, project.)

We can then define the probability of a history by

$$\Pr{}^{C+}_{\rho}(\alpha)=\int H^{C+}(\alpha)\cdot\rho \tag{16}$$

in the classical case and

$$\Pr{}^{C+}_{\rho}(\alpha)=\mathsf{Tr}(H^{C+}(\alpha)\cdot\rho) \tag{17}$$

in the quantum case.

The classical expression automatically determines a (generally stochastic) macrodynamics (that is, a probability measure over histories); the quantum expression does provided that all the coarse-grained distributions are diagonalised by projection onto the macrostates: that is, provided that

$$\mathcal{C}\cdot\rho=\sum_M P(M)\cdot\mathcal{C}\cdot\rho \tag{18}$$

where the sum ranges over macrostates. This condition is satisfied automatically by the measurement and correlation-discard rules (the latter rules, recall, build in the former); it will be satisfied by the coarse-grained exemplar rules provided the exemplars are chosen appropriately; it will be satisfied approximately by the smearing rules given that the smearing function is small on macroscopic scales.

Examples in physics where this process is used to generate a macrodynamics include:[3]

[3] It is of interest to note that all these examples — and indeed all the examples of which I am aware — use the correlation-discard coarse-graining rule or the coarse-grained exemplar rule. The other rules, so far as I know, are used in foundational discussions but not in practical applications — though I confess freely that I have made no systematic study to verify this.

Boltzmann's derivation of the H theorem Boltzmann's "proof" that a classical gas approached the Maxwell-Boltzmann distribution requires the "Stosszahlansatz" — the assumption that the momenta of gas molecules are uncorrelated with their positions. This assumption is in general very unlikely to be true (cf. the discussion in Sklar (1993, pp.224-7)), but we can reinterpret Boltzmann's derivation as the forward dynamics induced by the coarse-graining process of simply discarding those correlations.

More general attempts to derive the approach to equilibrium As was already noted, the kind of mathematics generally used to explore the approach of classical systems to equilibrium proceeds by partitioning phase space into cells and applying a smoothing process to each cell. (See Sklar (1993, pp. 212-4) for a discussion of such methods; I emphasise once again that at this stage of the discussion I make no defence of their conceptual motivation.)

Kinetic theory and the Boltzmann equation Pretty much all of non-equilibrium kinetic theory operates, much as in the case of the H theorem, by discarding the correlations between different particles' velocities. Methods of this kind are used in weakly interacting gases, as well as in the study of galactic dynamics (Binney and Tremaine 2008). The BBGKY hierarchy of successive improvements of the Boltzmann equation (cf Sklar (1993, pp. 207–210) and references therein) can be thought of as introducing successively more sophisticated coarse-grainings which preserve N-body correlations up to some finite N but not beyond.

Environment-induced decoherence and the master equation Crucially given our goal of understanding the asymmetry of quantum branching, quantitative results for environment-induced decoherence are generally derived by (in effect) alternating unitary (and entangling) interactions of system and environment with a coarse-graining defined by replacing the entangled state of system and environment with the product of their reduced states (derived for each system by tracing over the other system).

Local thermal equilibrium In pretty much all treatments of heat transport (in, for instance, oceans or stars) we proceed by breaking the system up into regions large enough to contain many particles, small enough to treat properties such as density or pressure as constant across them. We then take each system to be at instantaneous thermal equilibrium at each time, and study their interactions.

In most of the above examples, the coarse-graining process leads to deterministic macrodynamics. Some (rather theoretical) examples where it does not are:

Rolling dice We don't normally do an explicit simulation of the dynamics that justifies our allocation of probability 1/6 to each possible outcome of rolling a die. But qualitatively speaking, what is going on is that (i) symmetry considerations tell us that the region of phase space corresponding to initial conditions that lead to any given outcome has Liouville volume 1/6 of the total initial-condition volume; (ii) because the dynamics are highly random, any reasonably large and reasonably Liouville-smooth probability distribution over the initial conditions will therefore overlap to degree 1/6 with the region corresponding to each outcome; (iii) any coarse-graining process that delivers coarse-grained states which are reasonably large and reasonably Liouville-smooth will therefore have probability 1/6 of each outcome.

Local thermal equilibrium for a self-gravitating system Given a self-gravitating gas, the methods of local thermal equilibrium can be applied, but (at least in theory) we need to allow for the fact that a distribution which initially is fairly sharply peaked on a spatially uniform (and so, un-clumped) state will in due course evolve through gravitational clumping into a sum of distributions peaked on very non-uniform states. In this situation, the macrodynamics will be highly non-deterministic, andso if we want to coarse-grain by discarding long-range correlations, we first need to decompose the distribution into macroscopically definite components.

Decoherence of a system with significant system-environment energy transfer If we have a quantum system being decohered by its environment, and if there are state-dependent processes that will transfer energy between the system and environment, then macro-level correlations between, say, system centre-of-mass position and environment temperature may develop, and tracing these out will be inappropriate. Again, we need to decompose the system into components with fairly definite macroproperties before performing the partial trace.

4 Time reversibility in coarse-grained dynamics

The process used to define forward dynamics — as the name suggests — is explicitly time-asymmetric, and this makes it at least possible that the forward dynamics are themselves time-irreversible. In fact, that possibility is in general fully realised, as we shall see in this section.

Given a dynamical trajectory of the microdynamics, we know that we can obtain another dynamical trajectory by applying the time-reversal operator and then running it backwards. Following this, we will say that a given segment of a dynamical trajectory of the coarse-grained dynamics is time-reversible if the corresponding statement holds true. That is, if $\rho(t)$ is a segment of a dynamical trajectory (for $t \in [t_1, t_2]$) then it is reversible iff $T\rho(-t)$ is a segment of a dynamical trajectory (for $t \in [-t_2, -t_1]$).[4]

Although the microdynamics is time-reversible, in general the coarse graining process is not, and this tends to prevent the existence of time-reversible

[4] Note that I assume, tacitly, that the dynamics is time-translation-invariant, as is in fact the case in both classical and quantum systems in the absence of explicitly time-dependent external forces.

coarse-grained trajectories. It is, in fact, possible to define a function S_G — the *Gibbs entropy* — on distributions, such that S_G is preserved under microdynamical evolution and under time reversal, but such that for any distribution ρ, $S_G(\mathcal{C}\rho) \geq S_G(\rho)$, with equality only if $\mathcal{C}\rho = \rho$. (And so, since the forward dynamics consists of alternating microdynamical evolution and coarse-graining, S_G is non-decreasing on any dynamical trajectory of the forward dynamics.) In the classical case, we take

$$S_G(\rho) = -\int \rho \ln \rho \tag{19}$$

and in the quantum case we use

$$S_G(\rho) = -\mathsf{Tr}(\rho \ln \rho). \tag{20}$$

(At the risk of repetitiveness: I am assuming *absolutely nothing* about the connection or otherwise between this function and thermodynamic entropy; I use the term “entropy” purely to conform to standard usage.) Of the coarse-graining methods described above, the facts that correlation-discard, measurement, and smearing increase Gibbs entropy are well known results of (classical or quantum) information theory; the exemplar rule will increase Gibbs entropy provided that the exemplars are chosen to be maximal-entropy states, which we will require.

The existence of a Gibbs entropy function for $\mathcal{C}$ is not itself enough to entail the irreversibility of the $\mathcal{C}+$ dynamics. Some coarse-grained distributions might actually be carried by the microdynamics to other coarse-grained distributions, so that no further coarse-graining is actually required.

I will call a distribution *Boring* (over a given time period) if evolving its coarse-graining forward under the microdynamics for arbitrary times within that time period leads only to other coarse-grained distributions, and *Interesting* otherwise. The most well-known Boring distributions are stationary distributions — distributions whose forward time evolution under the microdynamics is themselves — such as the (classical or quantum) canonical and microcanonical distributions; any distribution whose coarse-graining is stationary is also Boring. On reasonably short timescales, generic states of many other systems — planetary motion, for instance — can be treated as Boring or nearly so.[5] However, if the ergodic hypothesis is true for a given system (an assumption which otherwise will play no part in this paper), then on sufficiently long timescales the only Boring distributions for that system are those whose coarse-grainings are uniform on each energy hypersurface.

If a segment of a dynamical trajectory of the $\mathcal{C}+$ dynamics contains any distributions that are Interesting on timescales short compared to the segment's length, that segment is irreversible. For in that case, nontrivial coarse-graining occurs at some point along the trajectory, and so the final Gibbs entropy is strictly greater than the initial Gibbs entropy. Time reversal leaves the Gibbs entropy invariant, so it follows that for the time-reversed trajectory, the initial

[5] More precisely, in general a system's evolution will be Boring on timescales short relative to its Lyapunov timescale.

Gibbs entropy is higher than the final Gibbs entropy. But we have seen that Gibbs entropy is nondecreasing along any dynamical trajectory of the forward dynamics, so the time-reversed trajectory cannot be dynamically allowed by those dynamics.

So: the coarse-graining process $\mathcal{C}$ takes a dynamical system (classical or quantum mechanics) which is time reversal invariant, and generates a new dynamical system ($\mathcal{C}+$, the forward dynamics induced by $\mathcal{C}$) which is irreversible. Where did the irreversibility come from? The answer is hopefully obvious: it was put in by hand. We could equally well have defined a *backward* dynamics induced by $\mathcal{C}$ ($\mathcal{C}$- for short) by running the process in reverse: starting with a distribution, coarse-graining it, evolving it backwards in time by some time interval, and iterating. And of course, the time reversal of any dynamical trajectory of $\mathcal{C}+$ will be a dynamical trajectory of $\mathcal{C}-$, and vice versa.

It follows that the forward and backwards dynamics in general make contradictory claims. If we start with a distribution at time t_i, evolve it forwards in time to t_f using the $\mathcal{C}+$ dynamics, and then evolve it backwards in time using the $\mathcal{C}-$ dynamics, in general we do *not* get back to where we started.

This concludes the purely mathematical account of irreversibility. One more physical observation is needed, though: the forward dynamics induced by coarse-graining classical or quantum mechanics has been massively empirically successful. Pretty much all of our quantitative theories of macroscopic dynamics rely on it, and those theories are in general *very* well confirmed by experiment. With a great deal of generality — and never mind the conceptual explanation as to *why* it works — if we want to work out quantitatively what a large physical system is going to do in the future, we do so by constructing a coarse-graining-induced forward dynamics.

On the other hand (of course), the *backwards* dynamics induced by basically any coarse-graining process is not empirically successful at all: in general it wildly contradicts our actual records of the past. And this is inevitable given the empirical success of the forward dynamics: on the assumption that the forward dynamics are not only predictively accurate now but also were in the past (a claim supported by very extensive amounts of evidence) then — since they are in conflict with the backwards dynamics — it cannot be the case that the backwards dynamics provides accurate ways of retrodicting the past. Rather, if we want to retrodict we do so via the usual methods of scientific inference: we make tentative guesses about the past, and test those guesses by evolving them forward via the forward dynamics and comparing them with observation. (The best-known and best-developed account of this practice is the Bayesian one: we place a credence function on possible past states, deduce how likely a given present state is conditional on each given past state,and then use this information to update the past-state credence function via Bayes' Theorem.)

5 Microdynamical underpinnings of the coarse-grained dynamics

In this section and the next, I turn my attention from the practice of physics to the justification of that practice. That is: given that (we assume) it is really the macrodynamics induced by the microdynamics — and not the coarse-grained dynamics — that describe the actual world, under what circumstances do those two processes give rise to the *same* macrodynamics?

There is a straightforward technical requirement which will ensure this: we need to require that for every history α,

$$\mathcal{C}H(\alpha)\rho = H^{C+}(\alpha)\rho. \tag{21}$$

That is, the result of alternately evolving ρ forward under the fine-grained dynamics and restricting it to a given term in a sequence of macro-properties must be the same, up to coarse-graining, as the result of doing the same with the coarse-grained dynamics. If ρ and $\mathcal{C}$ jointly satisfy this condition (for a given history space), we say that ρ is *forward predictable* by $\mathcal{C}$ on that history space. (Mention of a history space will often be left tacit.) Note that in the quantum case, if ρ is forward predictable by $\mathcal{C}$, it follows that the macrohistories are decoherent with respect to ρ.

I say "Forward" because we are using the coarse-grained *forward* dynamics. Pretty clearly, we can construct an equivalent notion of backwards predictability, using the backward coarse-grained dynamics and the anti-histories mentioned in section 2. And equally clearly, ρ is forward predictable by $\mathcal{C}$ if and only if its time reverse is backwards predictable by $\mathcal{C}$.

Forward predictability is closely related to the (slightly weaker) notion of *forward compatibility.* A distribution ρ is *forward compatible* with a given coarse-graining map $\mathcal{C}$ if evolving ρ forward under the microdynamics and then coarse-graining at the end gives the same result as evolving ρ forward (for the same length of time) under the coarse-grained dynamics. (Note that forward compatibility, unlike forward predictability, is not defined relative to any given history space.) Forward predictability implies forward compatibility (just consider the trivial history, where the macrostate at each time is the whole state space) and the converse is true in systems that are macrodeterministic. More generally, if $H(\alpha)\rho$ is forward compatible with $\mathcal{C}$ for all histories α in some history space, then ρ is forward predictable by $\mathcal{C}$ on that history space.

Prima facie, one way in which forward compatibility could hold is if the coarse-graining rule is actually physically implemented by the microdynamics: if, for instance, a distribution ρ is taken by the micrograined dynamics to the distribution $\mathcal{C}\rho$ on timescales short compared to those on which the macroproperties evolve, then all distributions will be forward compatible with $\mathcal{C}$. And indeed, if we want to explain how one coarse-grained dynamics can be compatible with another even coarser-grained dynamics, this is very promising. We can plausibly explain the coarse-graining rule for local equilibrium thermodynamics, for instance, if we start from the Boltzmann equation and deduce that systems satisfying that equation really do evolve quickly into distributions which are locally canonical. (Indeed, this is the usual defence given of local thermal equilibrium models in textbooks.)

But clearly, this cannot be the explanation of forward compatibility of the *fine-grained* dynamics with any coarse-graining rule. For by construction, the coarse-graining rules invariably increase Gibbs entropy, whereas the fine-grained dynamics leave it static. One very simple response, of course, would be just to postulate an explicit modification to the dynamics which enacts the coarse-graining. In classical mechanics, Ilya Prigogine has tried to introduce such modifications (see, e. g. , Prigogine (1984) and references therein); in quantum mechanics, of course, the introduction of an explicit, dynamical rule for the collapse of the wavefunction could be thought of as a coarse graining, and the final chapter of *Time and Chance* can be seen as developing this idea.

However, at present there remains no direct empirical evidence for any such dynamical coarse-graining. For this reason, I will continue to assume that the unmodified microdynamics (classical or quantum) should be taken as exact.

Nonetheless, it would not be surprising to find that distributions are, in general, forward compatible with coarse graining. Putting aside exemplar rules for coarse-graining, there are strong heuristic reasons to expect a given distribution generally to be forward compatible with the other three kinds of rules:

- A distribution will be forward compatible with a smearing coarse-graining rule whenever the microscopic details of the distribution do not affect the evolution of its overall spread across phase space. Whilst one can imagine distributions where the microscopic structure is very carefully chosen to evolve in some particular way contrary to the coarse-grained prediction, it seems heuristically reasonable to suppose that generically this will not be the case, and that distributions (especially reasonably widespread distributions) which differ only on very small lengthscales at one time will tend to differ only on very small lengthscales at later times. (However, I should note that I find this heuristic only *somewhat* plausible, and in light of the dearth of practical physics examples which use this rule, would be relaxed if readers are unpersuaded!)

- A distribution will be forward compatible with a correlation-discard coarse-graining rule whenever the details of the correlation do not affect the evolution of the macroscopic variables. Since macroscopic properties are typical local, and correlative information tends to be highly delocalised, heuristically one would expect that generally the details of the correlations are mostly irrelevant to the macroscopic properties — only in very special cases will they be arranged in just such a way as to lead to longer-term effects on the macroproperties.

- A distribution will be forward compatible with a measurement coarse-graining rule (which, recall, is nontrivial only for quantum theory) whenever interference between components of the distribution with different macroproperties does not affect the evolution of those macroproperties.

> This is to be expected whenever the macroproperties of the system at a given time leave a trace in the microproperties at that time which is not erased at subsequent times: when this is the case, constructive or destructive interference between branches of the wavefunction cannot occur. Decoherence theory tells us that this will very generically occur for macroscopic systems: particles interacting with the cosmic microwave background radiation or with the atmosphere leave a trace in either; the microscopic degrees of freedom of a non-harmonic vibrating solid record a trace of the macroscopic vibrations, and so forth. These traces generally become extremely delocalised, and are therefore not erasable by local physical processes. In principle one can imagine that eventually they re-localise and become erased — indeed, this will certainly happen (on absurdly long timescales) for spatially finite systems — but it seems heuristically reasonable to expect that on any realistic timescale (and for spatially infinite systems, perhaps on any timescale at all) the traces persist.

At least in the deterministic case, forward compatibility implies forward predictability; even in probabilistic cases, these kind of heuristics suggest — again, only heuristically — that forward predictability is generic.

In any case, my purpose in this paper is not to prove detailed dynamical hypotheses but to identify those hypotheses that we need. So — given the above heuristic arguments — we could try postulating a

Bold Dynamical Conjecture: For any system of interest to studies of irreversibility, all distributions are forward predictable by the appropriate coarse-grainings of that system on the appropriate history space for that system.

It is clear that, were the Bold Dynamical Conjecture correct, it would go a long way towards explaining why coarse-graining methods work.

But the line between boldness and stupidity is thin, and — alas — the Bold Dynamical Conjecture strides Boldly across it. For suppose $X = \mathcal{C}\rho$ is the initial state of some Interesting segment of a dynamical trajectory of the forward coarse-grained dynamics (Interesting so as to guarantee that Gibbs entropy increases on this trajectory) and that X' is the final state of that trajectory (say, after time t). Then by the Bold Dynamical Conjecture, X' can be obtained by evolving ρ forward for time t under the fine-grained dynamics (to some state ρ', say) and then coarse-graining.

Now suppose we take the time-reversal TX' of X' and evolve it forward for t seconds under the coarse-grained forward dynamics. By the Bold Dynamical Conjecture, the resultant state could be obtained by evolving $T\rho'$ forward for t seconds under the fine-grained dynamics and then coarse-graining. Since the fine-grained dynamics are time-reversible, this means that the resultant state is the coarse-graining of $T\rho$. And since coarse-graining and time reversal commute, this means it is just the time reverse TX of X.

But this yields a contradiction. For Gibbs entropy is invariant under time reversal, so $S_G(TX) = S_G(X)$ and $S_G(TX') = S_G(X')$. It is non-decreasing on any trajectory, so $S_G(TX) \geq S_G(TX')$. And it is increasing (since the trajectory is Interesting) between X and X', so $S_G(X') > S_G(X)$. So the Bold Dynamical Conjecture is false; and, more generally, we have shown that if $\mathcal{C}\rho$ is any coarse-grained distribution on a trajectory of the forward coarse-grained dynamics which has higher Gibbs entropy than the initial distribution on that trajectory, then $T\rho$ is *not* forward compatible with $\mathcal{C}$.

So much for the Bold Dynamical Conjecture. But just because not *all* distributions are forward compatible with $\mathcal{C}$, it does not follow that none are; it does not even follow that most aren't. Indeed, the (admittedly heuristic) arguments above certainly seem to suggest that distributions that are in some sense "generic" or "typical" or "non-conspiratorial" or somesuch term will be forward compatible with the coarse-grainings. In general, the only known way to construct *non*-forward compatible distributions is to evolve a distribution forward under the fine-grained dynamics and then time-reverse it.

This suggests a more modest proposal:

Simple Dynamical Conjecture (for a given system with coarse-graining $\mathcal{C}$): Any distribution whose structure is at all simple is forward predictable by $\mathcal{C}$;any distribution *not* so predictable is highly complicated and as such is not specifiable in any simple way *except* by stipulating that it is generated via evolving some other distribution in time (for instance, by starting with a simple distribution, evolving it forwards in time, and then time reversing it).

Of course, the notion of "simplicity" is hard to pin down precisely, and I will make no attempt to do so here. (If desired, the Simple Dynamical Conjecture can be taken as a family of conjectures, one for each reasonable precisification of "simple".) But for instance, any distribution specifiable in closed functional form (such as the microcanonical or canonical distributions, or any distribution uniform over a given (reasonably-simply-specified) macroproperty, would count as 'specifiable in a simple way'.

In fact, it will be helpful to define a *Simple* distribution as any distribution specifiable in a closed form in a simple way, without specifying it via the time evolution of some other distribution. Then the Simple Dynamical Conjecture is just the conjecture that all Simple distributions are forward predictable by the coarse-graining. Fairly clearly, for any precisification of the notion of Simple, a distribution will be Simple iff its time reverse is.

Are individual states (that is, classical single-system states or quantum pure states) Simple? It depends on the state in question. Most classical or quantum states are not Simple at all: they require a great deal of information to specify. But there are exceptions: some product states in quantum mechanics will be easily specifiable, for instance; so would states of a classical gas where all the particles are at rest at the points of a lattice. This in turn suggests that the Simple Dynamical Conjecture may well fail in certain classical systems (specifically, those whose macrodynamics is in general indeterministic): Simple classical systems will generally have highly unusual symmetry properties and so may behave anomalously. For example, a generic self-gravitating gas will evolve complex and highly asymmetric structure because small density fluctuations get magnified over time, but a gas with no density fluctuations whatever has symmetries which cannot be broken by the dynamics, and so will remain smooth at all times.

This appears to be an artefact of classical mechanics, however, which disappears when quantum effects are allowed for. A quantum system with a similar dynamics will evolve into a superposition of the various asymmetric structures; in general, the classical analogue of a localised quantum wavefunction is a narrow Gaussian distribution, not a phase-space point. So I will continue to assume that the Simple Dynamical Conjecture holds of those systems of physical interest to us.

6 Microdynamical origins of irreversibility: the classical case

It is high time to begin addressing the question of what all this has to do with the real world. I begin with the classical case, although of course the quantum case is ultimately more important. The question at hand is: on the assumption that classical microphysics is true for some given system, what additional assumptions need to be made about that system in order to ensure that its macroscopic behaviour is correctly predicted by the irreversible dynamics generated by coarse-graining?

The most tempting answer, of course, would be "none". It would be nice to find that absolutely any system has macroscopic behaviour well-described by the coarse-grained dynamics. But we know that this cannot be the case: the coarse-grained dynamics is irreversible, whereas the microdynamics is time-reversal-invariant, so it cannot be true that all microstates of a system evolve in accordance with the coarse-grained dynamics. (A worry of a rather different kind is that the coarse-grained dynamics is in general probabilistic, whereas the classical microdynamics are deterministic.)

This suggests that we need to supplement the microdynamics with some restrictions on the actual microstate of the system. At least for the moment, I will assume that such restrictions have a probabilistic character; I remain neutral for now as to how these probabilities should be understood.

A superficially tempting move is just to stipulate that the correct probability distribution over microstates of the system is at all times forward predictable by the coarse-graining. This would be sufficient to ensure the accuracy of the irreversible dynamics, but it is all but empty: to be forward predictable by the coarse graining *is* to evolve, up to coarse-graining, in accordance with the irreversible dynamics.

Given the Simple Dynamical Conjecture, an obvious alternative presents itself: stipulate that the correct probability distribution over microstates is at all times Simple. This condition has the advantage of being non-empty, but it suffers from two problems: it is excessive, and it is impossible. It is excessive because the probability distribution at one time suffices to fix the probability distribution at all other times, so there is no need to independently impose it at more than one time. And it is impossible because, as we have seen, in general the forward time evolution of a Simple distribution is not Simple. So if we're going to impose Simplicity as a condition, we'd better do it once at most.

That being the case, it's pretty clear when we have to impose it: at the beginning of the period of evolution in which we're interested. Imposing Simplicity at time t guarantees the accuracy of the forward coarse-grained dynamics at times later than t; but by time reversibility (since the time-reverse of a Simple distribution is Simple) it also guarantees the accuracy of the *backwards* coarse-grained dynamics at times earlier than t, which we need to avoid. So we have a classical recipe for the applicability of coarse-grained methods to classical systems: they will apply, over a given period, only if at the beginning of that period the probability of the system having a given microstate is specified by a Simple probability function.

So, exactly when should we impose the Simplicity criterion? There are basically two proposals in the literature:

1. We should impose it, on an ad hoc basis, at the beginning of any given process that we feel inclined to study.

2. We should impose it, once and for all, at the beginning of time.

The first proposal is primarily associated with the objective Bayesian approach pioneered by Jaynes (see, e. g., Jaynes (1957a, 1957b, 1968) — and I have to admit to finding it incomprehensible. In no particular order:

- We seem to be reasonably confident that irreversible thermodynamic processes take place even when we're not interested in them;

- Even if we are uninterested in the fact that our theories predict anti-thermodynamic behaviour of systems before some given time, they still do. (i.e., the problem that our theories predict anti-thermodynamic behaviour doesn't go away just because they make those predictions before the point at which we are "inclined to study" the system in question.)

- The direction of time is put in by hand, via an a priori assumption that we impose our probability measure at the beginning, rather than the end, of the period of interest to us. This seems to rule out any prospect of understanding (for instance) humans themselves as irreversible physical systems.

Perhaps the most charitable way to read the first proposal is as a form of strong operationalism, akin to the sort of operationalism proposed in the foundations of quantum mechanics by, e. g., Fuchs and Peres (2000). In this paper, though, I presuppose a more realist approach to science, and from that perspective the second proposal is the only one that seems viable: we must impose Simplicity at the beginning of time. The time asymmetry in irreversible processes is due to the asymmetry involved in imposing the condition at one end of time rather than the other.

(Incidentally, one can imagine a cosmology — classical or quantum — according to which there is no well-defined initial state — for instance, because the state can be specified at arbitrarily short times after the initial singularity but not at the singularity itself, or because the notion of spacetime itself breaks down as one goes further into the past. If this is the case, some somewhat more complicated formulation would presumably be needed, but it seems unlikely that the basic principles would be unchanged. For simplicity and definiteness, I will continue to refer to "the initial state".)

At this point, a technical issue should be noted. My definition of the Simple Dynamical Conjecture was relative to a choice of system and coarse-graining; what is the appropriate system if we want to impose Simplicity at the beginning of time? The answer, presumably, is that the system is the universe as a whole, and the coarse-graining rule is just the union of all the coarse-graining rules we wish to use for the various subsystems that develop at various times. Presumably there ought to exist a (probably imprecisely-defined) maximally fine-grained choice of coarse-graining rule such that the Simple Dynamical Conjecture holds for that rule; looking ahead to the quantum-mechanical context, this seems to be what Gell-Mann and Hartle (2007) mean when they talk about a maximal quasi-classical domain.

So: if the probabilities we assign to possible initial states of the Universe are given by a Simple probability distribution, and if we accept classical mechanics as correct, we would predict that the coarse-grained forward dynamics are approximately correct predictors of the probability of the later Universe having a given state. We are now in a position to state an assumption which suffices to ground the accuracy of the coarse-grained dynamics.

Simple Past Hypothesis (classical version): There is some Simple distribution ρ over the phase space of the Universe such that for any point x, $\rho(x)\delta V$ is the objective probability of the initial state of the Universe being in some small region δV around x.

(By "objective probability" I mean that the probabilities are not mere expressions of our ignorance, but are in some sense objectively correct.)

To sum up: if (a) the world is classical; (b) the Simple Dynamical Conjecture is true of its dynamics (for given coarse-graining $\mathcal{C}$); (c) the Simple Past Hypothesis is true, then the initial state of the world is forward predictable by the$\mathcal{C}+$ dynamics: the macrodynamics defined by the $\mathcal{C}+$ dynamics is the same as the macrodynamics induced by the microdynamics.

7 Microdynamical origins of irreversibility: the quantum case

Rather little of the reasoning above actually made use of features peculiar to classical physics. So the obvious strategy to take in the case of quantum mechanics is just to formulate a quantum-mechanical version of the Simple Past Hypothesis involving objective chances of different pure states, determined by some Simple probability distribution.

There are, however, two problems with this: one conceptual, one technical. The technical objection is that quantum distributions are density operators, and the relation between density operators and probability distributions over pure states is one-to-many. The conceptual objection is that quantum mechanics already incorporates objective chances, and it is inelegant, to say the least, to introduce additional such.

However, it may be that no such additional objective chances are in fact necessary, for two reasons.

1. There may be many pure states that are Simple and which are reasonable candidates for the state of the very early Universe.

2. It is not obvious that pure, rather than mixed, states are the correct way to represent the states of individual quantum systems.

To begin with the first: as I noted previously (p.20) there is no problem in quantum mechanics in regarding certain pure states as Simple, and the (as always, heuristic) motivations for the Simple Dynamical Conjecture are no less true for these states. As for the second, mathematically speaking mixed states do not seem obviously more alien than pure states as representations of quantum reality. Indeed, if we wish to speak at all of the states of individual systems in the presence of entanglement, the only option available is to represent them by mixed states. And since the universe appears to be open, and the vacuum state of the universe appears to be entangled on all lengthscales (cf. Redhead (1995) and references therein), even the entire observable universe cannot be regarded as in a pure state.

This being the case, I tentatively formulate the quantum version of the Simple Past Hypothesis as follows.

Simple Past Hypothesis (quantum version): The initial quantum state of the Universe is Simple.

What is the status of the Simple Past Hypothesis? One way to think of it is as a hypothesis about whatever law of physics (fundamental or derived) specifies the state of the very early universe: that that law requires a Simple initial state. Indeed, if one assumes that probabilistic physical laws must be simple (which seems to be part of any reasonable concept of 'law'), and that simplicity entails Simplicity, all the Simple Past Hypothesis amounts to is the

Past Law Hypothesis: The initial quantum state of the Universe is determined by some law of physics.

Alternatively, we might think of the Simple Past Hypothesis as a (not very specific) conjecture about the contingent facts about the initial state of the Universe, unmediated by law. Indeed, it is not clear that there is any very important difference between these two readings of the Hypothesis. In either case, the route by which we come to accept the Hypothesis is the same: because of its power to explain the present-day observed phenomena, and in particular the success of irreversible macrodynamical laws. And on at least some understandings of 'law' (in particular, on a Humean account like that of Lewis (1986) where laws supervene on the actual history of the Universe) there is not much metaphysical gap between (i) the claim that the initial state of the Universe has particular Simple form X and this cannot be further explained, and (ii) the claim that it is a law that the initial state of the Universe is X.

8 A low entropy past?

The suggestion, espoused by Albert, that the origin of irreversibility lies in constraints on the state of the early universe is hardly new: it dates back to Boltzmann, and has been espoused in recent work by, among others, Penrose (1989, 2004), Goldstein (2001), and Price (1996). But their Past Hypotheses differ from mine in an interesting way. Mine is essentially a constraint on the microstate of the early universe which is essentially silent on its macrostate (on the assumption that for any given macroscopic state of the universe, there is a Reasonable probability distribution concentrated on that macrostate). But the normal hypothesis about the past is instead a constraint on the macrostate of the early universe:

Low Entropy Past Hypothesis: The initial macrostate of the universe has very low thermodynamic entropy.

Is such a Hypothesis needed in addition to the Simple Past Hypothesis? I think not. For if the Simple Past Hypothesis is true (and if the Simple Dynamical Conjecture is correct) then it follows from the Hypothesis and our best theories of microdynamics that the kind of irreversible dynamical theories we are interested in — in particular, those irreversible theories which entail that thermodynamic entropy reliably increases — that the entropy of the early universe was at most no higher than that of the present universe, and was therefore "low" by comparison to the range of entropies of possible states (since there are a great many states with thermodynamic entropy far higher than that of the present-day universe). So the Low Entropy Past "Hypothesis" is not a Hypothesis at all, but a straightforward prediction of our best macrophysics — and thus, indirectly, of our best microphysics combined with the Simple Past Hypothesis.

It will be helpful to expand on this a bit. On the assumption that the relevant irreversible dynamics (in this case, non-equilibrium thermodynamics) is predictively accurate, predictions about the future can be made just by taking the current state of the universe and evolving it forward under those dynamics. Since the dynamics do not allow retrodiction, our route to obtain information about the past must (as noted earlier) be more indirect: we need to form hypotheses about past states and test those hypotheses by evolving them forward and comparing them with the present state. In particular, the hypothesis that the early universe was in a certain sharply specified way very hot, very dense, very uniform, and very much smaller than the current universe — and therefore much lower in entropy than the current universe[6] — does very well under this method: conditional on that hypothesis, we would expect the current universe to be pretty much the way it in fact is. On the other hand, other hypotheses — notably the hypothesis that the early universe was much higher in entropy than the present-day universe — entail that the present-day universe is fantastically unlikely, and so very conventional scientific reasoning tells us that these hypotheses should be rejected.

In turn, we can derive the assumption that our irreversible dynamical theories are predictively accurate by assuming (i) that our microdynamical theories are predictively accurate, and (ii) that the Simple Past Hypothesis and the Simple Dynamical Conjecture are true. So these hypotheses jointly give us good reason to infer that the early universe had the character we believe it to have had. On the other hand, (i) alone does not give us reason to accept (ii). Rather, we believe (ii) because combined with (i), it explains a great deal of empirical data — specifically, the success of irreversible dynamical theories.

The difference between the Simple Past Hypothesis and the Low Entropy Past Hypothesis, then, does not lie in the general nature of our reasons for believing them: both are epistemically justified as inferences by virtue of their explanatory power. The difference is that the Reasonable Past Hypothesis, but not the Low Entropy Past Hypothesis, is justified by its ability to explain the success of thermodynamics (and other irreversible processes) *in general.* The Low Entropy Past Hypothesis, by contrast, is justified by its ability to explain *specific features* of our current world. (Although the hypothesis that does this is better understood as a specific cosmological hypothesis about the state of the early universe, rather than the very general hypothesis that its entropy was low.)

Albert himself gives a particularly clear statement of his framework for inducing the (Low Entropy) Past Hypothesis, which makes an interesting contrast to my own. He makes three assumptions:

1. That our best theory of microdynamics (which for simplicity he pretends is classical mechanics) is correct.

2. That the Low Entropy Past Hypothesis is correct.

3. That the correct probability distribution to use over current microstates is the uniform one, conditionalised on whatever information we know (notably, the Low Entropy Past Hypothesis).

He also makes a tacit mathematical conjecture, which is a special case of the Simple Dynamical Conjecture: in my terminology, he assumes that those distri-

[6] It is widely held that (i) such a universe ought to be much *higher* in entropy than the present-day universe, but (ii) this supposed paradox is solved when gravity is taken into account. This is very confused; I attempt to dispel the confusion in Wallace (2009).

butions which are uniform over some given macrostate and zero elsewhere are forward compatible with coarse-graining.

Now, (2) and (3) together entail that the correct distribution to use over initial states (and Albert is fairly explicit that "correct" means something like "objective-chance-giving") is the uniform distribution over whatever particular low entropy macrostate is picked out by the Low Entropy Past Hypothesis. Since these distributions are Simple, Albert's two assumptions entail the Simple Past Hypothesis. But the converse is not true: there are many Simple distributions which are not of the form Albert requires, but which (given the Simple Dynamical Conjecture) are just as capable of grounding the observed accuracy of irreversible macrodynamics.

Put another way: let us make the following abbreviations.

SPH: Simple Past Hypothesis

LEPH: Low Entropy Past Hypothesis

UPH: Uniform Past Hypothesis: the hypothesis that the initial distribution of the universe was a uniform distribution over some macrostate

SDC: Simple Dynamical Conjecture

PAμ: Predictive Accuracy of Microphysics (i. e., our current best theory of microphysics is predictively accurate)

PAM: Predictive Accuracy of Macrophysics (i. e., the macrodynamics derived from microphysics by coarse-graining is predictively accurate)

My argument is that

$$SPH + SDC + PA\mu \longrightarrow PAM. \tag{22}$$

Albert's (on my reading) is that

$$LEPH + UPH + SDC + PA\mu \longrightarrow PAM. \tag{23}$$

But in fact

$$UPH \to SPH \tag{24}$$

so actually $LEPH$ appears to play no important role in Albert's argument. All that really matters is that the initial distribution was uniform over some macrostate; the fact that this macrostate was lower entropy than the present macrostate is then a straightforward inference from PAM and the present-day data.

9 Conclusion

There are extremely good reasons to think that, in general and over timescales relevant to the actual universe, the process of evolving a distribution forward under the microdynamics of the universe commutes with various processes of coarse-graining, in which the distribution is replaced by one in which certain fine structures — most notably the small-scale correlations and entanglements between spatially distant subsystems — are erased. The process of alternately coarse-graining in this manner and evolving a distribution forwards leads to dynamical processes which are irreversible: for instance, when probabilistic, they will have a branching structure; where a local thermodynamic entropy is definable, that entropy will increase. Since coarse-graining, in general, commutes with the microdynamics, in general we have good grounds to expect distributions to evolve under the microdynamics in a way which gives rise to irreversible macrodynamics, at least over realistic timescales.

Given that the microdynamics is invariant under time reversal, if this claim is true then so is its time reverse, so we have good reason to expect that, in general, the evolution of a distribution both forward and backwards in time leads to irreversible macrodynamics on realistic timescales. It follows that the claim can be true only 'in general' and not for *all* distributions, since — for instance — the time-evolution of a distribution which does behave this way cannot in general behave this way. However, we have no reason to expect this anomalous behaviour except for distributions with extremely carefully chosen fine-scale structure (notably those generated from other distributions by evolving them forwards in time). I take this to be a more accurate expression of Goldstein's idea of 'typicality': it is not that systems are guaranteed *to achieve equilibrium* unless they or their dynamics are "ridiculously special"; it is that only in "ridiculously special" cases will the micro-evolution of a distribution not commute with coarse-graining. Whether, and how fast, a system approaches thermal equilibrium is then something that can be determined via these coarse-grained dynamics.

In particular, it seems reasonable to make the Simple Dynamical Conjecture that reasonably simple distributions do not show anomalous behaviour. If the correct distribution for the Universe at some time t is simple in this way, we would expect that macrophysical processes after t are well-described by the macrodynamics generated by coarse-graining (and so exhibit increases in thermodynamic entropy, dispersal of quantum coherence, etc), in accord with the abundant empirical evidence that these macrodynamics are correct. But we would also expect that macrophysical processes *before* t are not at all described by these macrodynamics — are described, in fact, by the time reversal of these macrodynamics — in wild conflict with the empirical evidence. But if t is the first instant of time (or at least, is very early in time) then no such conflict will arise.

It follows that any stipulation of the boundary conditions of the Universe according to which the initial distribution of the Universe is reasonably simple will (together with our microphysics) entail the correctness of our macrophysics. Since any law of physics specifying the initial distribution will (essentially by the nature of a law) require that initial distribution to be reasonably simple, it follows that any law which specifies the initial distribution suffices to ground irreversible macrodynamics.

It is virtually tautologous that if microscopic physics has no time asymmetry but the emergent macroscopic dynamics does have a time asymmetry, that time asymmetry must be due to an asymmetry in the initial conditions of the universe. The most common proposal for this asymmetry is the proposal that the initial distribution is the uniform distribution over a low-entropy macrostate. From the point of view of explaining irreversibility, all the work in this proposal is being done by the "uniform distribution" part: the low-entropy part alone is neither necessary nor sufficient to establish the correctness of the irreversible macrodynamics, though of course if the initial macrostate is a maximum-entropy state then its macroevolution will be very dull and contradicted by our observations.

And in fact, the only special thing about the uniformity requirement is that we have good (if heuristic) grounds to expect the microdynamical evolution of uniform distributions to be compatible with coarse-grainings. But we have equally good (if equally heuristic) grounds to expect this of any simply specified distribution. So really, the asymmetry of the Universe's macroscopic dynamics is not a product of the particular form of the physical principle which specifies the initial conditions of the Universe: it is simply a product of some such principle being imposed at one end of the Universe rather than at the other.

Acknowledgements

I'd like to take the opportunity to thank David Albert for many stimulating discussions over the last ten years: from no-one with whom I disagree even half as much have I learned even half as much.

I'd also like to thank Jeremy Butterfield and Simon Saunders for helpful comments on earlier versions of this paper, and to acknowledge valuable conversations with Harvey Brown, Wayne Myrvold, Roman Frigg, and Jos Uffink.

References

Albert, D. Z. (2000). *Time and Chance.* Cambridge, MA: Harvard University Press.

Binney, J. and S. Tremaine (2008). *Galactic Dynamics* (2nd ed.). Princeton: Princeton University Press.

Callender, C. (2009). The past hypothesis meets gravity. In G. Ernst and A. Hütteman (Eds.), *Time, Chance and Reduction: Philosophical Aspects of Statistical Mechanics*, Cambridge. Cambridge University Press. Available online at http://philsci-archive.pitt.edu/archive/00004261.

Earman, J. (2006). The 'past hypothesis': Not even false. *Studies in the History and Philosophy of Modern Physics 37*, 399–430.

Frigg, R. (2008). Typicality and the approach to equilibrium in Boltzmannian statistical mechanics. Available online at http://philsci-archive.pitt.edu.

Fuchs, C. and A. Peres (2000). Quantum theory needs no "interpretation". *Physics Today 53*(3), 70–71.

Gell-Mann, M. and J. B. Hartle (2007). Quasiclassical coarse graining and thermodynamic entropy. *Physical Review A 76*, 022104.

Goldstein, S. (2001). Boltzmann's approach to statistical mechanics. In J. Bricmont, D. Dürr, M. Galavotti, F. Petruccione, and N. Zanghi (Eds.), *In: Chance in Physics: Foundations and Perspectives*, Berlin, pp. 39. Springer. Available online at http://arxiv.org/abs/cond-mat/0105242.

Home, D. and M. A. B. Whitaker (1997). A conceptual analysis of quantum Zeno: Paradox, measurement and experiment. *Annals of Physics 258*, 237–285.

Jaynes, E. (1957a). Information theory and statistical mechanics. *Physical Review 106*, 620.

Jaynes, E. (1957b). Information theory and statistical mechanics ii. *Physical Review 108*, 171.

Jaynes, E. (1968). Prior probabilities. *IEEE Transactions on Systems Science and Cybernetics SSC-4*, 227.

Lebowitz, J. (2007). From time-symmetric microscopic dynamics to time-asymmetric macroscopic behavior: An overview. Available online at http://arxiv.org/abs/0709.0724.

Lewis, D. (1986). *Philosophical Papers, Vol. II.* Oxford: Oxford University Press.

Misra, B. and E. C. G. Sudarshan (1977). The Zeno's paradox in quantum theory. *Journal of Mathematical Physics 18*, 756.

Penrose, R. (1989). *The Emperor's New Mind: concerning computers, brains and the laws of physics.* Oxford: Oxford University Press.

Penrose, R. (2004). *The Road to Reality: a Complete Guide to the Laws of the Universe.* London: Jonathon Cape.

Price, H. (1996). *Time's Arrow and Archimedes' Point.* Oxford: Oxford University Press.

Prigogine, I. (1984). *Order out of Chaos.* Bantam Books.

Redhead, M. (1995). More ado about nothing. *Foundations of Physics 25*(1), 123–139.

Saunders, S., J. Barrett, A. Kent, and D. Wallace (Eds.) (2010). *Many Worlds? Everett, Quantum Theory, and Reality*, Oxford. Oxford University Press.

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

Wallace, D. (2009). Gravity, entropy, and cosmology: in search of clarity. Forthcoming.

Wallace, D. (2011). *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation.* Oxford: Oxford University Press.