The Nature of the Past Hypothesis

David Wallace

(This is a lightly edited and referenced transcript of the author's talk at the Philosophy of Cosmology conference, Tenerife, September 2014)

There's a narrative about the nature of asymmetry in time which can be caricatured like this. Firstly, there is our fundamental physics - if we're being really careful about it, this is the standard model or our preferred post-standard-model physics; in reality, in the practical cases we think about, it's more likely to be ordinary quantum mechanics or maybe even classical Hamiltonian dynamics. In any case, it's supposed to be the physics of the micro-constituents of the world. And it is time-reversal invariant and shows no particular direction of time.

And secondly there is the observed world, which is full of various kinds of observed asymmetries: dynamical asymmetries, entropic asymmetries, causal, psychological asymmetries, and so on. And the general way we set the problem up is as a contradiction between what our physics says, which is that the world is time-reversal invariant, and what we see around us, which isn't time-reversal invariant.

I want to suggest the advantages of a slightly more nuanced way of thinking about the problem. It's really not the case that all of physics, or even most of physics, or even very much of physics, frankly, is "fundamental" physics. In the middle – between the fundamental physics at the bottom, and the directly observed macro-world at the top – we have a huge range of what we might call higher-level (or "emergent") dynamical systems, governed by higher-level dynamical equations. I'm thinking of the equations of fluid dynamics; I'm thinking of the Boltzmann equation that governs dilute gases and many similar systems; I'm thinking of Langevin equation and the Fokker-Planck equation that govern Brownian motion; I'm thinking of the equations of radioactive decay. (In principle I'm also thinking about all the various equations and rules of the higher sciences, but we can confine our attention here just to the panoply of different systems and different equations that we study in physics.)

Actual physics is a plethora of different dynamical systems governed by different sorts of laws. And if you look at those higher-level laws you find – not universally, but very frequently – that *they* have a whole range of properties which aren't shared by the fundamental physics. I want to focus

on two particular properties like this.

Firstly, they tend to have a lack of time symmetry. It's worth pausing on what that means for a second, because in one sense the Standard Model (and its plausible successors) don't have time symmetry either: they're symmetric under CPT but not under T alone. But higher-level theories are time-reversal-noninvariant in a more important way that the Standard Model doesn't share: they're generally *irreversible*, which is to say that their dynamical maps are many-to-one, either in the literal sense that they take different initial states to the *same* later state or in the sense that they take different initial states to final states that get closer and closer together over time, so that to any given grain of resolution they might as well be taken as the same final state. And frequently these higher-level theories are not just irreversible in general, but have attractors: particular points in their state space such that all the states which share the same conserved quantities as that attractor end up at that attractor, according to the equations of those theories.

And secondly – somewhat less importantly for my purposes, but not irrelevantly – the dynamical equations of these higher-level theories tend to be probabilistic. Which is to say: sometimes they're stochastic equations; sometimes they're equations for the evolution of classical probability distributions; quantum-mechanically they tend to be equations for the evolution of mixed states. In general we tend to recover determinism for these kinds of theories only in a law-of-large-numbers sense, and only when we're talking about systems with a lot of degrees of freedom. Something like the Boltzmann equation will do as an example; characteristically[1] that's set out probabilistically as an evolution equation for a one-particle marginal (whether that's a density operator, as in quantum theory, or a classical probability distribution), but the multi-body correlations are weak enough and the particle numbers are large enough that at the end of the process we can treat the predictions as deterministic. That's not always how it goes - Boltzmann famously derived the classical Boltzmann equation directly without going through a probability route[2] – but it tends to be the general pattern.

Pause for a second – and set aside for a second how any of this higher-

[1] See, e.g., R. Balescu, *Statistical Dynamics: Matter Out of Equilibrium* (London: Imperial College Press, 1997).
[2] See H.R.Brown, W.Myrvold and J.Uffink, "Boltzmann's H-theorem, its discontents, and the birth of statistical mechanics", *Stud.Hist.Phil.Mod.Phys* 40 (2009) pp. 174-191, and references therein.

level physics is linked to the underlying fundamental physics – and ask just how we'd go about making claims about the past and the future of a system governed by these kinds of laws. In a theory governed by *reversible* dynamical laws; there's going to be a fairly obvious symmetry in how we do it. How do you use the theory to learn about the future? You look at what the present state is, turn the handle of the dynamics, and out comes a prediction about the future state that can be checked against experiment. How do you use the theory to learn about the past? Pretty much the same. You plug in the present state, you run the dynamical equations backwards, that tells you what the past state is supposed to be, and then you compare it with what it actually was. And there's a reason we call this by the neologism "*retrodiction*": it's to suggest that we're doing the same kind of thing as a *pre*diction about the future.

In a theory governed by *irreversible* dynamical laws, prediction is going to be a similar kind of game: you plug in the present state of the world, evolve it forward in time, and out comes a statement – possibly probabilistic – about the future state of the world. We're not in a position to retrodict in the same way, because in an irreversible theory, the present state is typically *not* going to determine a past state. That could be because the dynamics is deterministic but irreversible, so that many past states are compatible with the present state; it could also be because the dynamics is probabilistic and just isn't really in the business of telling us *anything* about what the past looked like.

What we do in practice is a kind of "guess and check". At the crudest level, we take a guess as to what looks like a reasonable past state, we evolve it forward and see what the state *would* be now, we compare with what it actually *is*, and we iterate that process. If we want to be slightly more careful, more systematic, more formal about that, we could put something like a Bayesian prior distribution over our collective initial states, use that Bayesian prior to work out a probability distribution over possible present states, conditionalize on the actual present state and see where that leads us. However this process is made precise, let's call it *historical inference*, and distinguish it sharply from retrodiction. It's our normal means by which we learn about the past in these kind of theories.

(It's worth reminding ourselves of the cases where we actually use retrodiction, just to see the contrast with historical inference. If we want to work out the dates of some eclipse that's mentioned in some fifth century BC history we really do carry out pretty much the time-reverse of the calculations that we use to work out the dates of an eclipse in the next century – that is, we really do just run the equations of the solar system

backwards. But this is very much the exception which proves the rule, and even then it only applies approximately.)

In a world which was, hypothetically, really governed by these kind of irreversible (and often probabilistic) laws, it wouldn't be particularly mysterious that we had a whole bunch of psychological asymmetries, causal asymmetries, entropic asymmetries, and so forth. If our underlying physics contained a whole bunch of irreversibility and violation of time-reversal invariance, we'd expect that the observed world would also have those features.

All this suggests a different way of setting up the problem of asymmetry in time from the way it's normally set up, the way we started with. In that approach, typically we say: how do we reconcile the fundamental-physics level directly with the observed world? My suggestion, as a friendly amendment to this way of thinking, is that there are advantages in keeping the discussion internal to the equations of physics, and asking: how can we reconcile the bottom level, the fundamental physics level, with the emergent higher levels, the levels of irreversible dynamical equations? If we can sort this out, the remaining step – reconciling that higher level with the observed-macro-world asymmetries – looks rather tractable.

Why do I make this suggestion? There are three reasons. The first is just division of labour. It's quite a job to say: ok, how am I to reconcile the microscopic physics of the world with the fact that I remember the past but not the future? There are so many different layers, and so many different bits of science, filtering into that kind of story that trying to do the whole thing all in one go inevitably involves radically simplifying models, and dynamical assumptions that we're not remotely in a position to check. On the other hand, if we can outsource the problem of understanding why we remember the past but not the future to those with relevant expertise in memory and cognition and evolution and so forth while handing them on a plate a whole bunch of underlying physics that has time irreversibility in it already, then the task looks a bit more tractable than trying to do the whole thing ourselves.

The second reason is that we need to understand the fundamental/emergent relation anyway. An account of why we remember the past and not the future, or an account of why ice melts in qualitative terms, that doesn't also tell us quantitatively how it is that we can have the Langevin equation or something similar holding compatibly with our micro-physics hasn't finished the job, because that equation is demonstrably correct for some physical systems and we need to understand why.

Conversely, if we can account for the latter, if we can get a grip on how to reconcile higher-level irreversible dynamical systems with bottom-level reversible dynamical systems, we'll get a lot of the rest more or less for free.

The third reason, and the one I want to dwell on, is: there's something funny about saying it's a deep mystery how we can reconcile time reversible microphysics with time irreversible higher-level physics. In one sense it *is* a mystery: we can prove a whole bunch of formal incompatibilities and show that no time-asymmetric physics can be derived from a time-symmetric starting point. But on the other hand, mostly we don't get our higher-level dynamical physics purely from phenomenology, purely from experiment. To a very large extent we actually do derive it from the microphysics – or perhaps, to avoid begging the question, we at least *construct* it from the microphysics. We have a collection of thoroughly used and highly successful techniques for starting with microlevel physics and getting out equations governing the higher level physics. In fact the great bulk of what we call the evidence for our low-level physics is actually mediated through this kind of process. And we don't just get the qualitative form of the higher-level equations out, we actually get the coefficients.

For instance, think about the decay rates of particles. Those are governed by a time-irreversible decay equation, but we get that equation out from quantum field theory. And we get out the decay coefficient while doing it. So at some level we clearly know how to do this, or at least we have a trick that very reliably works. And which works, as philosophers would say, *projectably*, in two senses. It allows us to work out dynamical equations which seem to be laws for the systems in question, in that they apply to those systems wherever they are in space and time. And the fact that we can do this also turns out to be projectable: if we take a novel physical system where we haven't yet tried to work out what the macro dynamics are, and we use these kinds of techniques, we tend to get out empirically correct laws.

So that suggests that if we want to understand where the time asymmetry comes into physics, we ought to be looking at what ingredient we are actually putting in when we construct the Langevin equation, or the radioactive decay equation, or any of these higher-level equations. And indeed, this is a sanity check on extant claims of where the asymmetry is. If, for instance, it’s claimed that the origin of temporal asymmetry is a specific low entropy boundary condition for the Universe, we ought to be able to see how it is that this low entropy boundary condition, perhaps

indirectly, underpins whatever is actually being put in to our derivational process to get out the Langevin equation and the like from our microphysics.

And that moves me on to the second half of what I want to talk about: let's actually have a look in a little bit more detail at these derivations, and see what's going on.[3]

I'm going to be very qualitative, and as general as I can manage, but of course the activity of deriving higher level dynamical equations from lower level dynamical equations is enormously wide and varied – indeed, in a sense it's the great bulk of physics – and I only know small corners of it. So let's stipulate: I'm talking about a certain subclass of such processes. That subclass is not empty; conjecturally I'd say that that subclass contains an awful lot of what we do, but it's not particularly my brief to say that it covers every low-level/high-level derivation known.

Here's a generic model of how I'm going to think about things. To a large extent what we're doing when we construct higher level physics is some kind of coarse-graining. At the kinematic level, there's a state space S_L of the low-level theory, and there's a state-space S_H of the high-level theory, and there's what we might call a *reduction map* that takes us from points of S_L to points of S_H. That map is typically going to be many-to-one: it's going to take a whole group of low-level states and associate them with a single high-level state.

The paradigmatic example of this is something like the way we do fluid or gas dynamics. Here the low-level theory, at least classically, is going to be the Hamiltonian dynamics of 10^{23} point particles interacting under some force laws and the states in the high-level theory are going to be specified by giving the pressure and density and velocity of the gas averaged over, say, one-cubic-micron cells. That means that S_H is still a large dimensional state space, but it's a lot smaller than the low-level space S_L: A whole bunch of different particle arrangements are going to be associated to a single fluid states. And it's going to follow, of course, just from that reduction map, that if I take a trajectory in S_L defined by the low-level dynamics, that trajectory is going to be mapped to some path in S_H.

If we move outside the particular case I've just discussed and ask generally how we might try to set up high-level to low-level

[3] For technical details here, see D. Wallace, *The Emergent Multiverse: Quantum Theory According to the Everett Interpretation* (Oxford: Oxford University Press, 2012), chapter 9, and references therein.

correspondences in physics, there's a slightly unfortunate tendency at this point to go into pragmatics and epistemology. Sometimes you'll hear people say that what we're doing when we go from the low-level to the high-level theory is: we're keeping only those features of the system that we're interested in and discarding those that don't interest us. For instance, maybe we're not interested in the precise positions of the gas, we just want to know its bulk state.

That isn't going to work. I'm not actually very interested in the cubic micron of gas over there in that corner; I'm not really very interested in the gas in this room at all. I could perfectly happily go through the rest of my life knowing nothing about it, and I imagine so could the rest of you. But it's still true that its dynamics is governed by the equations of fluid dynamics; that generality is no less true because I don't care about it.

Conversely, I'm actually quite interested in what the stock market looks like tomorrow. I would love to be able to do a dynamical reduction where I course-grained over everything except the leading numbers in the price index tomorrow and kept those. But however fascinated I might be about that I can't do it, not in a way that's predictively useful. What's going on here is that we're not really asking: when is the high-level theory defined by the degrees of freedom we're interested in? We're actually asking: when is the high-level theory defined by degrees of freedom for which we can write down autonomous dynamical rules?

What do I mean by that? The low-level theory has dynamics, which determine low-level trajectories, and each low-level trajectory determines a high-level trajectory, but there's no *a priori* guarantee that the low-level *dynamics* determines a high-level *dynamics.* There's no guarantee, for instance, that there can't be two trajectories in S_L whose images in S_H are identical up till some moment of time and then diverse. Which is to say that there's no guarantee that in moving from the low level to the high level I haven't discarded some information which is actually necessary to predict the future evolution of the high-level states.

So what we actually want are reduction maps that don't have this feature: that actually do generate a high-level dynamics from a low-level dynamics. In this situation we have something we can call a *dynamical reduction* process. Another way to put it is that we can find a dynamics on S_H such that evolving a state in S_L forward in time under the lower-level dynamics and then mapping it to S_H, or mapping it to S_H immediately and evolving it under the dynamics on S_H, gives the same results. In mathematical terminology, dynamical evolution, and the coarse-graining reduction map,

commute.

There doesn't *need* to be any such high-level dynamics, but often there is. And in particular, often that high-level dynamics is pretty robust against the fine-grained details of how we define the reduction map. (I talked about my gas on cubic micron cells, but clearly I could have chosen ten cubic microns or half a cubic micron.)

The existence of these high-level dynamical laws is, or ought to be, kind of surprising. After all, step back from the fact that we know it works empirically, and ask how is it that we can find equations for just a small number of degrees of freedom in a huge system that are autonomous, that can be studied dynamically whilst discarding the information at the other degrees of freedom? As with the stock market case, it's not generally true that that can be done: generally we can't just pick a subset of degrees of freedom and work out what they do, given that they're dynamically coupled to all the other degrees of freedom.

So why do high-level laws ever exist? Very, very broadly, we can see two reasons. Firstly, sometimes it happens because there really is a dynamical decoupling of some degrees of freedom from others, and generally that happens when we've got a symmetry of some kind in play. So it really is the case that, under appropriate approximations, the centre of mass degree of freedom of the Earth decouples from the zillions of other degrees of freedom of the Earth. And so I really can write down an evolution equation for the centre of mass of the Earth treated as a point with just three degrees of freedom. Even that's not perfect: if the gravitational field varies quickly on scales comparable to the length of the Earth, I'm going to have trouble, but to a first approximation I can do it. And I can do it basically because the symmetry structure of the dynamics lets me decouple the centre-of-mass degrees of freedom from the rest.

Much more commonly, as for instance in the gas, what's going on is not that there's complete decoupling of this kind: it's that the residual degrees of freedom, those discarded when we apply a coarse-graining reduction map, are very large in number and very random in the fine details of their dynamics, and each one of them is contributing only to a very small extent to the overall dynamics. So we can do a statistical trick: rather than keep track of them in bulk we can just keep track of their statistical averages; we can sum all their contributions up and treat the whole thing as a sort of generalized noise term. (And you see this in some of the ways one actually derives these kind of equations in detail.) Using that method is implicitly taking a bet that actually those fine-grain details don't matter,

and that we really can treat them as a sort of averaged-over noise.

That sounds great, but there's a sense in which we know that it can't really be true. And the sense is the following: if it really were the case that some reduction map from low-level to high-level state space was compatible with a completely accurate, robust reduced dynamics for the high-level theory, then if the dynamics of the low-level theory is time-reversal invariant the dynamics of the high-level theory had better be time-reversal invariant as well. And it isn't. So something went wrong.

And if we look at what went wrong – if we look at the mathematics of what we're doing here – what's going wrong is that it's not true that *any* distribution of the microscopic degrees of freedom with such and such average will behave in such and such a way. There will be ways of tuning and setting up the microscopic degrees of freedom, so that they're aligned in just the right way to break our assumptions about how the dynamics is going to work.

Let me give an example here – partly to help see how general this discussion is, it's not a statistical mechanical example in the usual sense. Think about radioactive decay. We have simple higher-level dynamics for decay: the probability of an unstable particle not decaying is exponential in time, and there are also terms for particles being kicked into undecayed states by absorbing the sort of particles that comprise the decay products.

Now ask how all this works quantum-mechanically. If I take a particle on my desk which has in fact not decayed and evolve it forward in time, it will evolve into a superposition of the undecayed particle and a whole bunch of decay products at different times. If I try evolving it backwards under Schrodinger's equation, I'll erroneously predict the time reverse of a decay: after all, the equation has a time reversal symmetry. How do I reconcile that with the fact that the particle has just been sitting on my desk undecayed?

The answer is that the full quantum state at the present time contains not just a term describing the undecayed particle but all the terms corresponding to the particle having decayed at various times in the *past*. And properly running the dynamics backwards means allowing for all of these other terms – which all have just the right mod-squared-amplitudes, and just the right phases, that they continually cancel out the decay terms that are produced from the backward evolution of the undecayed term. It's

because of that setup being aligned just right – in Everettian terms,[4] it's because of all the branches going backwards in time interfering in just the right way – that I get the wrong answer if I try to retrodict using the normal radioactive decay equations.

And for the same reason, if I take the time reverse of the current state (including all the decay terms as well as the undecayed term corresponding to our current experience), that state's *forward* time evolution won't match the normal prediction of radioactive decay. More or less any old generic state of the way the fields could be wouldn't do this but this very carefully prepared one, prepared by time evolving the system forward and then time-reversing, is set up in just the right way to do it.

This state is a *counter-example* state, a state for which our methods of statistical averaging fail to generate the right higher-level evolution. And in general, there will exist counter-example states whenever we try to derive irreversible higher-level dynamics from reversible lower-level dynamics.

What can we say in general about the counter-example states? Just that they're going to be very delicately and carefully structured. It's tempting to say that they're very low probability states, in some sense, but I want to warn against that, because we're mostly here talking about theories that are already formulated as probability theories or theories of mixed states, so our space of states here is a space of probability distributions or mixed states anyway. And talk of "low-probability probability distributions" is at least *prima facie* not well defined. So it's not that the counter-example states are low probability, exactly, but that heuristically, and in some cases demonstrably, they're going to be very delicate, complicated, carefully specified states. (Indeed, the only way we really know to write down any such states is the one we used in the case of radiation: take a simple state, evolve it forwards through time, and then time-reverse it.)

So then the question of why high-level dynamics of this kind in general works (and in particular, the question of why non-equilibrium statistical mechanics works) is going to be the question of why we're allowed to assume that initial state isn't like that, isn't one of the counter-example states. And of course if we take any given physical system – some small, mundane system, like the glass of water on my desk – it's not surprising that its initial state is not like that: its initial state has been prepared by a

[4] Everett, H. "Relative State Formulation of Quantum Mechanics", *Rev.Mod.Phys.* 29 (1957), pp. 454-462. For further discussion see D. Wallace, *The Emergent Multiverse: Quantum Theory according to the Everett Interpretation* (Oxford: Oxford University Press, 2012).

whole bunch of other dynamical processes, so unless those dynamical processes are really, really carefully set up in a certain way, or unless the initial state that we feed into *those* processes was very carefully set up in a certain way, we'd be very surprised to find that the initial state of our mundane system was one of the very delicately assembled states which count as the exceptions to the usual reduction rule.

What we've done there is push back the requirement that the state is not careful and special in this sense back to an earlier system: the system which led dynamically to our mundane system having the initial state it had. And of course – and this is my justification in talking about this material in a conference on cosmology – if you keep pushing these things back and back, following from our initial system, to the system that generated it, to the system that generated *that* system, … where you're going to get to is a requirement that the initial state of the Universe as a whole is not one of these very special counter-example states.

To sum up: what we seem to get out if we look at the actual structure of reduction and emergence in physics is a claim about the initial state of the universe – something that's hardly novel in discussions of time asymmetry – but a claim that is slightly different from the usual claims.[5] We're not requiring anything of the *macro*state of the early universe: in particular, we're not requiring it to have low entropy. Instead, we're requiring something about its *micro*structure: we're requiring it *not* to be set up in the kind of delicately correlated ways that invalidate our general averaging moves.

From this perspective, far from introducing a condition which makes the initial universe very, very special (the usual ways "Past-Hypothesis" claims about the early Universe are phrased), we're asking that it should *not* be very special. We need to be a bit careful, though, because of course if we suppose that the universe is closed and has a final state as well as an initial state, and if we take this "not very special" initial state and run it forward, then – precisely because this licenses us to apply higher-level, emergent dynamics – the *final* state is going to be really, really special. It's going to have built into it just the right correlations such that if time-reversed and run backwards it *won't* be governed by the macro level

[5] Here I have in mind the neo-Boltzmannian approach espoused by, e.g., D. Albert (*Time and Chance*; Cambridge, MA: Harvard University Press, 1999), R. Penrose (*The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics;* Oxford: Oxford University Press, 1989), S. Carroll (*From Eternity to Here: The Quest for the Ultimate Theory of Time;* New York: One World, 2010), S. Goldstein ("Boltzmann's approach to statistical mechanics", in J. Bricmont, D. Durr, M. Galavotti, F. Petruccione, and N. Zanghi (eds.), *Chance in Physics: Foundations and Perspectives*, Berlin: Springer, 2001; available online at http://arxiv.org/abs/cond-mat/0105242) or H. Price (*Time's Arrow and Archimedes' Point*; Oxford: Oxford University Press, 1996).

dynamics, but rather by their time reverse. So we haven't somehow dissolved away the need for an asymmetric assumption in. But that assumption looks a bit different, conceptually and mathematically, from what we're used to in these discussions: we're not supposing that the state at one end of time is particularly low-entropy, but that it's relatively simple, relatively free of complicated and delicate correlative structure.

(What *should* we say about the fact that the early universe has a very low entropy? Well, it certainly seems to need explanation, but it doesn't seem to be very different from other facts about the initial condition of the universe. How do we learn about the fact that the early universe has low entropy? By historical inference: we take our possible guesses as to what the initial state is, we evolve them forward, we compare with what we have. And since our macro-level dynamical theories are entropy-increasing, given that they are correct the fact that the early universe has a low entropy compared to the present day universe is a claim that is not very difficult to get that out historically.)

So: my suggested way of proceeding gives us, in David Albert's terms,[6] a different sort of transcendental assumption that we need to make in order that our physics works. But it's not a transcendental assumption of the kind we can actually empirically check in cosmology; it's a transcendental assumption about the fine-grained micro-level delicate structure of the early universe. And it's basically the assumption that it hasn't got much of it.

Question and Answer Session

David Albert: Thank you first as usual for a beautiful talk. I guess I have two small comments. One about the issue of division of labour that you were talking about. Here's the thing: presumably what we're interested in showing, among the asymmetries between past and future. For example, we're interested in showing not merely that epistemic access to the past is different than epistemic access to the future for human beings or for mammals, or for terrestrial organisms, or biological organisms in general. It's something much more general and much more fundamental than that. So that in that sense there is a sort of natural expectation that it should have some fairly direct link to the fundamental physics. If we were to encounter Martians that had time-reversed epistemic access relative to us we would be flabbergasted. So actually there's a sense in which it doesn't feel like a job for neurologists or psychologists or experts in human

[6] D. Albert, *Time and Chance* (Cambridge, MA: Harvard University, Press, 1999), ch. 6.

cognition. There's a sense in which it feels like a job for the fundamental physics. That's one thing.

A related point: The people who talk about features of the initial macro-state of the universe are trying to do a job slightly different from the one you describe here. That is, they're not just trying to justify the macro-dynamics or to explain the macro-dynamics. What such people are usually trying to do is something in a way more ambitious. I mean, maybe that's a foolhardy task, but they're trying to *both* justify the macro-dynamics *and* sort of systematize the whole process of inference towards past and future. So, for example, on the way of reasoning that you're describing, there are a bunch of earlier states that historical inference could lead to. One is the state we think pertained five minutes ago, the macro-state of the world. Another is the macro-state of the world we think pertained ten minutes ago. And so on and so forth, all the way back to the Big Bang. So I think I completely agree with you that if the task at hand is just to explain the macro-dynamics you need much less than that. If the task you're looking to do is more ambitious than that, you may need more.

David Wallace: It's absolutely right of course that we want a very general explanation of the epistemic asymmetries, and I agree with you it's a job for physics, but it's not necessarily a job for *fundamental* physics. We live in a world where the degrees of freedom are the macroscopic functionals are governed by time-asymmetric dynamical laws. In a situation like that it's unsurprising to find these kinds of general epistemic asymmetries; there's further work to try and understand them and get them out but it's not a mystery in that sense.

Having said that, in some sense this is spoils to the victor: if we can give that explanation directly in terms of fundamental physics, why not?

As to the other point: I don't disagree there at all. But it's sometimes said that the past hypothesis is specifically something we need to explain the fact that statistical mechanics techniques work. My claim is: that's not really what's doing the work. Put in your framework, what's actually doing all the work is the probability distribution over the initial macro condition, not the choice of macro-condition.

Carlo Rovelli: If I had to cover both talks, David and David [Albert] this morning, my comment would be that according to my own understanding everything that has been said is exactly right, as far as I understand. (Which, of course, might be due to the fact that I'm a theoretical physicist, I don't get saddled with philosophy.) But it seems to me: didn't we know

that; Isn't that the way we understand things; Haven't you just put clearly what was understood? And in fact wasn't it all understood by Boltzmann, especially in what he wrote *after* he got all the criticisms to his H theorem.

I want to make a point related to that. First, let me say that I talked about the problem of the special initial condition in my first talk this morning; perhaps my talk should have been *after* yours because the point of departure of my talk was the conclusion that you arrived at. But what I want to say is that after getting the criticism for his H theorem Boltzmann got to this point, which I think is not often appreciated: he was thinking in probabilistic terms, he was thinking in terms of equilibrium fluctuations, and what he proved – or rather, what he stated was the case, and in fact what has been proved recently by people in statistical theory, is the following. That if you take a statistical system and you look at the fluctuations of its entropy (which of course is not going to maximal, it's going to fluctuate) you can ask why, if I'm at a certain point away from the maximum, I find that it goes up in one direction in time and down in one direction in time, while my theory tells me that it goes down in both directions. And here's an answer to that: it's a beautiful answer. Given a value the most probable situation is that it's a peak of a fluctuation. So this makes it clear that his own result is valid not because it breaks time-reversal invariance; it confirms time-reversal invariance and it shows that given a situation in which you're out of equilibrium the most likely solution is going to be in *both* directions. And of course [in the context of our observations] the only way to make sense of that is to go back to the origin of the universe, so making it a cosmological posit. The theorem has been proven quite recently in statistical mechanics rigorously; Boltzmann just guessed it.

David Wallace: I want to pick up on the first point. The job of the philosopher of physics is to tell people the physics they know and then say that it's profound. I'm actually almost serious about that, not just being self-deprecating: a lot of the point of this kind of work is to take things that are tacitly grasped by people, and and understand clearly and consistently and explicitly how those things should be understood. It almost shades in to pedagogy at some level.

But I do think there's a degree of mismatch between what is said in general and verbal terms about why systems equilibrate, and what you actually find if you go down into the weeds and look at the way we construct and derive our detailed quantitative understanding of equilibration. I think that's of some interest, and in particular I think it's somewhat striking that the role of low entropy assumptions is more

indirect and much less clear than it can seem in the general discussions.

I've been interested in this stuff partly because I thought: okay, I buy the general claim that the way to understand why statistical mechanics works is because the initial universe had a low entropy, so let's actually have a look at how those equations work and see where the low entropy assumption in particular is coming in. And then it turned out to be more complicated than that.

Simon Saunders: I think the value of your talk, David, for me anyway, is that it seems to undercut a sense of mystery or some sense of strangeness or weirdness about how special the initial condition has to be. It seems – and Penrose has made a lot of this[7] – that by virtue of the fact that the final state of the universe , given black holes and eventually black hole evaporation, has a fabulously higher entropy than it has now, one gets to the picture that the early state of the universe had to amazingly, precisely fine tuned so as to have low entropy. And your punchline, in a way is: no, that it's not that special.

Don Page: That's very controversial, I didn't think you [Wallace] said that. Did you really say that?

Simon Saunders: Let me carry on, because I'm presenting a gloss on what David is saying and maybe David will tell me I've got the gloss wrong. My gloss is that the initial state of the universe, far from being special, precisely *isn't* at the microscopic level carefully calibrated so as to have the sorts of coincidental relations that could bring together what would look to us like time-reversed macroscopic phenomenology.

And I just want to push that, in that if you imagine Penrose, or some other physicist or mathematician, doing some other calculation which shows how yet more extraordinarily special the initial state seems to have to be, because of anthropic considerations, perhaps the right way to think of that is that what they have really shown is how extraordinarily high the final entropy of the universe can be, how extraordinarily large entropy can grow through physical processes. And perhaps black hole thermodynamics is an example of that, that prior to the understanding of black holes we didn't realize how large the entropy could get to be.

So I wonder if you would embrace that way of thinking about it. I think the comeback on that would be to say, well, okay, the initial state of the

[7] See R.Penrose, *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics* (Oxford: Oxford University Press, 1989), chapter 7.

universe has to be amazingly non-special at the microscopic level, it must not have all of these careful, finely tweaked relations among the, whatever it is, particles and so-forth, and you might say that what we're really learning is: you think you've got a non-special state because it doesn't have those finely-tuned correlations, well, guess what, it's got to be even more non-special than we thought it had to be.

Just to summarize then, suppose we learn of some new mechanism as we did in black hole thermodynamics, such that the final state of the universe could be even more extraordinarily high entropy, is that the right way to be surprised? Or is it Penrose's way: no, the surprise is to find how extraordinarily fine tuned the state of the universe has to be so as to be low entropy. Or, the third way that I'm offering as well, is it that we learn how extraordinarily non-special the initial state has to be?

David Wallace: I think this partly says what we mean by 'special'. So (and I take this to be what's worrying Don), I'm obviously *not* claiming that the initial macro state does not occupy an extremely small region of phase space. Clearly it does. What I'm claiming is that this fact about the macrostate is not the thing which is doing the explanatory work of saying why the law-like regularities of high-level statistical mechanics hold.

Now of course if we imagine a fictional world that (unlike in real cosmology) actually had an equilibrium state - a box-like world of some kind – then if it's initial state was unspecial in the macro sense it would be at equilibrium. The claim that the macro-dynamical laws held would then be boring because they just say: you're at the attractor, stay there. But nonetheless the thing that's doing the work in explaining *why* those (boring) law-like regularities hold is the non-specialness of the micro-structure within that macro-state. And you could perfectly well imagine a universe that was actually at equilibrium but whose micro-state *was* extremely special in just the right way that it went away from equilibrium. (Take a system that has just equilibrated, and time-reverse its microstate.)

On the broader question of the specialness of the initial *macro*state: I have to confess I'm not entirely clear, and I'm less clear the more I think about it, quite what is that's concerning about it. Granted, the theory that God created the universe by picking a point uniformly at random with respect to the Liouville measure on phase space is falsified by the data. But that wasn't a very plausible creation myth in the first place.

I think what's going on is something like this. In general, a really good way of studying a large system in a given macrostate is to assume its

microstate is chosen at random with respect to the Liouville measure over that macrostate. And we have reasonable dynamical grounds to explain why that's a good thing to do. But we're in danger of extrapolating this back to a more transcendental principle that it's *a priori* the case that a system is equally likely to be anywhere in phase space. And that's what gets you into skeptical catastrophes, where we say that somehow our memory of the past is completely unreliable because it's much, much more likely we've fluctuated in from a higher entropy state. But I don't think that there are any good reasons on the table to think that the right *a priori* probability distribution is uniform across phase space. There pausibly *are* good general *a priori* assumptions, coming from general epistemology or general philosophy of science, for thinking that the initial macro-state should be a relatively simple, or relatively easily describable state, but in gravitational systems that simplistic criterion tends to pick out low entropy states rather than high entropy states.[8]

So in conclusion, I'm not entirely sure what the fuss is about, so I'm not quite sure therefore what I should be saying in response to your trilemma at the end.

Cormac O'Rafferty: Many thanks to both Davids for talks which were extremely clear – that's not easy when you're going from discipline to discipline. If I could ask in that spirit a very, very simple question which for the philosophers is probably kindergarten. What do you make of Lee Smolin's view that there isn't *necessarily* a tension between the fact that some of our equations in particle physics don't include time, and the way that we see time in the observable universe. His simple answer to that [tension] is that since the Dirac equation physicists have been haunted by the notion that every solution *has* to represent the real world, whereas in fact mathematics is simply a representation of nature. That we fall into the old problem of confusing reality itself with our representation of reality? What do you make of people who duck the whole question by saying "there isn't necessarily a conflict there, this is simply a facet of the way we represent nature"?

David Wallace: The move of saying "our physics is not fully representing the world" is always available, but I think the only really good way to tell if our physics *can't* represent the world is to try really, really hard to represent the world and see if we fail. I'm a bit nervous about *assuming* that it's the case. That's a general philosophical nervousness about Smolin's move.

[8] For discussion of the conceptual features entropy in self-gravitating systems see D. Wallace, "Gravity, Entropy and Cosmology: In Search of Clarity", *Brit. J. Phil. Sci.* 61 (2010) pp. 513-540, and references therein.

Let me use that as a sort of advert for some of the virtues of the approach in my talk. The question of how we reconcile time-symmetric micro-dynamics with the panoply of phenomena the observed universe is so general, has so many different facets, that there's all sorts of space for more philosophical attempts to make the problem go away. The question of how it is that we can derive the Fokker-Planck equation from Newtonian dynamics, given that Newtonian dynamics is time-reversal symmetric and the Fokker-Planck equation isn't, and that the Fokker-Planck equation is probabilistic and Newtonian mechanics isn't; that question is much tighter and sharper. The problem is not some kind of philosophical paradox; it's a straight mathematical contradiction, so some of the assumptions in it must be wrong, and you can push at where those are.