What is Gibbs' Canonical Distribution?*

Kevin Davey

Abstract

Although the canonical distribution is one the central tools of statistical mechanics, the reason for its effectiveness is poorly understood. This is due in part to the fact that there is no clear consensus on what it *means* to use the canonical distribution to describe a system in equilibrium with a heat bath. In this paper, I examine some traditional views as to what sort of thing we should take the canonical distribution to represent. I argue that a less explored alternative, according to which the canonical distribution represents a time ensemble of sorts, has a number of advantages that rival interpretations lack.

1 Introduction.

One striking thing about the machinery of modern statistical mechanics is how effective it is. It is able to explain and predict a great number of facts about systems in equilibrium, and even some facts about systems out of equilibrium. More striking, however, is the fact that the *reason* for the effectiveness of this machinery is so poorly understood, despite so much effort by physicists, mathematicians and philosophers on this very issue.

Let us focus for now on the use of the canonical distribution to calculate quantities associated with a system in equilibrium with a heat bath. What justifies this procedure? Answering this question requires carefully untangling several distinct issues. First, we must ask what it *means* to describe the state of a system using the canonical distribution. Armed with the answer to this question, we must argue that we are *right* to describe systems in equilibrium with heat baths using the canonical distribution. Finally, we must argue that when we do correctly describe a system using the canonical ensemble, we are *justified* in expecting the value of a macroscopic observable on the system to be that given by integrating the observable over phase space using the canonical measure.

Some small progress can be made towards justifying the use of the canonical ensemble by carefully distinguishing these questions, and treating them one by

*Thanks to Brandon Fogel, Nick Huggett and John Norton for their feedback and suggestions.

one. In what follows, I provide a sketch of how to begin this task, focusing mainly on the question of what it means to describe the state of a system using the canonical distribution.

2 What is the Canonical Ensemble?

2.1 Equilibrium

We want to ask what it means to describe a system in equilibrium with a heat bath using the canonical distribution. It will be useful to begin by asking what it means to be *in equilibrium* with a heat bath at all. (For now, we take the concept of a heat bath as unproblematic.)

The concept of equilibrium has a disparate set of meanings, and we need to fix on one. Sometimes when we talk about a system being in equilibrium, we mean that the system is in a particular state – generally one of the macrostates described by the mathematical formalism of classical equilibrium thermodynamics – at a given point in time. I call this a *static* notion of equilibrium, insofar as it is generally used to determine whether a system is in equilibrium (possibly with something else) *at a given time*. This is not the only way in which we can think of equilibrium. In fact, when we talk about a gas being in equilibrium with a heat bath, we generally do not have a static conception of equilibrium in mind. The gas will exchange energy with the heat bath in such a way that it will occassionally find itself *out* of static equilibrium. So the following rough definition will suffice for our purposes: a system is in *equilibrium* with a heat bath if, as it evolves over time, it spends most of its time in the macrostate (or macrostates) associated with static equilibrium, only occasionally fluctuating into non-equilibrium macrostates. This is a temporally extended notion of equilibrium, insofar as saying that a system is in equilibrium with a heat bath in this sense is now making a claim about the behavior of the system *over time*.

Insofar as heat baths are supposed to be infinite, it may well turn out that any finite system placed in contact with a heat bath will behave in the way just described, if studied over a sufficiently large time scale. Thus, it may well turn out that any finite system in contact with a heat bath is in equilibrium with the heat bath, in the sense just defined. This should cause no confusion, so long as it is understood that there are different concepts of equilibrium, and a system in one sort of equilibrium need not be in another type of equilibrium.

2.2 Traditional Interpretations.

Let us return to the canonical distribution. According to Gibbs, the state of a system at equilibrium with a heat bath at temperature T is properly described by the 'canonical' distribution:

$$\rho(X,t) = \frac{e^{-\beta H(X)}}{\mathcal{Z}} \tag{1}$$

where X is a point in phase space, H is the Hamiltonian and $\mathcal{Z}$ the partition function. (For simplicity, we suppose that the Hamiltonian is time-independent, and that our system is constrained to lie in some finite volume of physical space.) But what does it *mean* to say of a system that it may be described by the canonical distribution (1)? I shall take it for granted that describing systems in contact with heat baths using the canonical distribution involves making some sort of probabilistic or statistical claim about such systems. But precisely what sort of probabilistic or statistical claim could this be?

One approach is to think of the canonical distribution as giving us *epistemic* probabilities. In its simplest form, the claim here is that that when we know a system is in equilibrium with a heat bath, it is rational to assign the degree of belief $\rho(X)dX$ to the proposition that the system is in region dX at time t, where ρ is the canonical distribution. But a claim like this is surely implausible. I might have a strong belief that the number of particles in a gas lies in a certain range (say, between 1.1×10^{23} to 1.2×10^{23} particles), but for each integer N in this range I will presumably assign a very low probability to N being the *exact* number of particles in the system. The canonical distribution, however, will assign probability 1 to the claim that the gas has exactly N particles, where N is the actual number of particles described in the Hamiltonian. In this case, the canonical distribution gives a probability of 1 where my epistemic probability is almost 0. A more sophisticated proponent of the epistemic interpretation of statistical mechanics might have something to say about this sort of worry,[1] though whether related worries arise is a question I am happy to leave open. For the purpose of this paper, I will focus instead on the idea that describing a system with the canonical distribution involves making an objective, non-epistemic probabilistic or statistical claim about the system at hand.

More specifically, according to the approaches that I do want to consider, describing a system with the canonical distribution involves making two observations: first, that the system is a member of a particular ensemble (or 'set' of systems), and second, that it is appropriate (in some sense to be specified) to describe this ensemble using the formula (1). To pursue this approach, we must first construct an ensemble of systems, and then explain in what sense it is appropriate to describe this ensemble with the formula (1).

There are a few ways to carve up possible ways of proceeding. One thing we might do is try to construct our ensemble out of actually occurring systems with Hamiltonian H, in contact with a heat bath at a given temperature T. We call this an *actualist* approach:

Actualist Approach: The ensemble should consist of some appropri-

[1] One might try to argue that our epistemic probabilities are given by *weighted sums* of canonical distributions, each involving a slightly different Hamiltonian, and that because the observables that matter have (roughly) the same expected value when calculated with this weighted sum as when calculated with the canonical distribution involving the *actual* Hamiltonian, we may take the canonical distribution to represent our epistemic state well enough for the purpose of calculating the expected macrostate of the system.

ately chosen set of actual (i.e., physically instantiated) systems with the Hamiltonian H, in contact with a heat bath at temperature T.

Given an ensemble constructed in accordance with this requirement, we might then claim that the statistics of the ensemble be (approximately) given by the canonical distribution. More precisely, we might claim that the proportion of systems in the ensemble occupying a fixed small region of phase space is (approximately) equal to the value of (1) in that region of phase space.[2] We call this an *(approximate) identity* requirement:

(Approximate) Identity Requirement: The proportion of systems in the ensemble in a fixed small region of phase space is (approximately) equal to the value of the canonical distribution in that region of phase space.

But the combination of an actualist approach with an approximate identity requirement is not promising at all. Presumably, there are systems that have only found themselves in contact with heat baths a small number of times in the history of the universe. Only a tiny region of their phase space will ever have been instantiated, and perhaps only a tiny region of their phase space will ever be instantiated. The canonical distribution will typically *not* describe the distribution of any set of instantiations of such systems.

In response to this, let us stick with the actualist approach, but try to relax the approximate identity requirement. To motivate our replacement for the approximate identity requirement, consider a coin that has been thrown 100 times. Take the hypothesis:

Hypothesis: If the coin were to be thrown continually, it would come up heads about half the time, and tails about half the time.

Suppose that the coin has come up heads 2 times and tails 98 times. In such a case, we would be inclined to say that *Hypothesis* has been strongly disconfirmed – where by 'strong disconfirmation' I mean (loosely speaking) a type of disconfirmation that is especially severe. If the coin came up heads 35 times and tails 65 times, we might say only that the hypothesis has been disconfirmed, but not strongly so. If the coin came up heads 48 times and tails 52 times, we would be inclined to say that *Hypothesis* has not been disconfirmed at all. Exactly how statistical confirmation or disconfirmation works, and whether and to what extent it needs to draw on a notion of prior probability, will not be important to us. The important point is the scientific commonplace that claims about the distribution of an ensemble can, in certain cases at least, be strongly disconfirmed by claims about a sample.

Let us suppose now that we have some procedure for sampling systems with Hamiltonian H, in contact with a heat bath at temperature T. The result is a (finite) set of systems distributed through phase space in some particular way. Consider now the hypothesis:

[2]We assume that the regions of phase space in question are sufficiently small that the value of (1) does not change significantly in them.

> *Hypothesis*: If we continually sample systems with Hamiltonian H in contact with a heat bath at temperature T, we will get an ensemble with the property that the proportion of systems in the ensemble in a small region of phase space is (approximately) equal to the value of (1) in that region of phase space.

This hypothesis is capable of being disconfirmed – and even strongly disconfirmed – by an actual (finite) sample of systems with Hamiltonian H in contact with a heat bath at temperature T. So instead of the approximate identity requirement, we impose the following requirement:

> Non-Disconfirmation Requirement: *Hypothesis* is not strongly disconfirmed by our actual (finite) sample of systems with Hamiltonian H in contact with a heat bath at temperature T.

The Non-Disconfirmation requirement can handle the case of uninstantiated or rarely instantiated systems much better, insofar as we typically do not take hypotheses about distributions to be strongly disconfirmable by very small (or even empty) samples.

The Non-Disconfirmation requirement is very modest – perhaps too modest to really be useful. It is interesting, then, that even a requirement this modest gets us into trouble. To see how, let us consider the details of the process by which we sample systems. Suppose we sample the microstate of a system with Hamiltonian H in contact with a heat bath at temperature T only at the moment the system is first placed in contact with the heat bath. The problem is that many systems are way out of static equilibrium when first placed in contact with a heat bath. Because of this, the method of sampling just described may consistently produce a disproportionate number of systems in microstates that are extremely rare according to the canonical distribution (1). As such, the hypothesis that continual sampling will result in an ensemble whose statistics are given by the canonical distribution will be strongly disconfirmed.

Perhaps it is unfair to sample the microstates of systems only at the moment they are placed in contact with heat baths. But if we just include *all* of the microstates of the system while it is in contact with a heat bath, the out-of-static-equilibrium microstates may still be disproportionately represented. (Imagine for instance a system with a very long relaxation time, such that up until now, such systems in contact with heat baths at a given temperature have spent almost all of their lives in regions of phase space that are extremely rare according to the canonical distribution.) It also will not do to sample the microstate the moment the system reaches static equilibrium, as the canonical distribution predicts energy fluctuations that will not be present if we only sample microstates with fixed energy. The absence of such energy fluctuations will count as strong disconfirmation of *Hypothesis*, particularly in cases in which the occurrence of such fluctuations is decently probable.

The most promising strategy is to sample the microstate of the system in contact with the heat bath only after some prespecified amount of time – some sort

of 'relaxation time' – appropriate to the system at hand. But I am skeptical as to whether this sort of approach can work. Highly contingent facts about the world, including facts about the way in which we choose to prepare systems, have a way of showing themselves in any such ensemble. For instance, consider a system that, at a given temperature, has multiple macroscopically distinguishable equilibrium states. Any system with potentially variable anisotropic structure, such as a crystal, will do. For a more artificial example, consider a 1-dimensional Ising ring with no external magnetic field, i.e., a system of N particles with Hamiltonian:

$$H = -J \sum_{i=1}^{N} \sigma_i \sigma_{i+1},$$

in which each σ_i takes on the value ± 1, and $\sigma_{N+1} = \sigma_1$. At low temperatures, this system organizes itself into large, macroscopic blocks of 1s or -1s. Because the Hamiltonian is unchanged under the transformation $\sigma \rightarrow -\sigma$, for each equilibrium macrostate at a given temperature there is a distinct equilibrium macrostate at the same temperature in which each macroscopic chunk is replaced by one of opposite sign. At sufficiently low temperatures, these equilibrium macrostates are clearly distinguishable.To make the argument that follows as clean as possible, let us suppose that systems with the given Hamiltonian do not occur naturally, but are always manufactured by us. It is perfectly possible for us, having manufactured such a system, to place it in contact with a heat bath at a certain cold temperature T much lower than that of the environment in which it is produced, if and only if the systems is in a *particular* equilbrium macrostate – for instance, iff the largest macroscopic block consists of 1s, rather than -1s. When placed in contact with the heat bath, such a system is very likely to 'relax' into a particular sort of macrostate – in this case, a macrostate in which it remains the case that the largest macroscopic block consists of 1s rather than -1s. As a result, insofar as we only sample systems after this relaxation time, our sample will consist primarily of microstates in a very specific region of phase space – namely, the region of phase space corresponding to this particular sort of macrostate. This ensemble will look extraordinarily unlikely, given the canonical probability distribution. If we place 1000 such systems in contact with heat baths, and essentially all of them end up with the largest macroscopic block consisting of 1s rather than -1s, it will be as if a coin thrown 1000 times came up heads essentially each time, which is a situation with extraordinarily low probability. In this case, we would have to say that *Hypothesis* is strongly disconfirmed.[3]

[3]This is not to deny that if any such system remained in contact with the heat bath long enough, there would eventually be large energy fluctuations, after which the system might find itself into one of the other possible macrostates at the given temperature. But this will typically occur only after an astronomically long time. So if we try to get around the problem by insisting that we only sample systems after sufficiently large energy fluctuations have occured, *all* our ensembles will typically be empty, and our description of these ensembles with the canonical distribution starts to border on vacuity. Whether obstacles like this can be satistfactorily surmounted is something I am happy to leave

I think that difficulties like these show that the entire actualist approach is in error. Insofar as our interpretation of the canonical distribution involves identifying an ensemble, I suspect that the ensemble will have to be populated not just with actual physical systems, but also with counterfactual systems. This is in agreement with Gibbs, who denied that his ensembles had any sort of '*objective existence*', saying instead that they were '*creations of the imagination*'. (See p.188 of [8].) But what counterfactual ensemble should we pick?

The traditional view, championed by Gibbs, was to simply *posit* an imaginary collection of systems distributed in accordance with the canonical distribution. According to this approach, when one describes a system using the canonical distribution, one is making a claim about an imaginary ensemble of which the particular system is a member. There are many well known problems with this approach. For instance, the actual system is a member of all sorts of imaginary ensembles – so why the privilege the canonical ensemble when it comes to making predictions?[4] I think the fact that problems like this are so difficult suggests that this way of thinking of the canonical distribution does not quite get things right, and that it makes sense to look for a more 'naturally occuring' counterfactual ensemble with which to work.

2.3 The Time-Ensemble

One particular proposal seems particularly natural. The rough idea is this – take a system in contact with a heat bath, and imagine allowing it to stay in contact with the heat bath indefinitely. Look at the distribution ψ of the set of microstates through which the system then passes. To describe a system in equilibrium with a heat bath using the canonical distribution is just to say that ψ *is* the canonical distribution.

Although I think this idea is basically right, it must be formulated very carefully in order to avoid being clearly wrong. The most obvious way to define a distribution ϕ on the set of microstates through which our system passes is as follows: focus at first on the microstates through which the system passes in a fixed interval of time from $t = 0$ to $t = T$. Let dX be a very small region of space space – for simplicity, we assume it is a ball centered around a point X of phase space – and let $\tau(dX)$ be the amount of time that the system spends in dX in the interval $[0, T]$. Then we would like to have:

$$\psi(X)|dX| \approx \frac{\tau(dX)}{T}$$

With this in mind, we define:

$$\psi(X) = \lim_{|dX|\to 0} \frac{\tau(dX)}{T|dX|}.$$

open – at any rate, these obstacles are challenging enough that it is worth considering alternative points of view.

[4]For extremely thorough discussions of this problem, though mainly with respect to the microcanonical rather than canonical ensemble, see [4] and Chapter 5 of [10].

where we imagine the radius of the open ball dX going to 0. The problem is that it may be shown to follow from this that $\psi(X) = 0$ almost everywhere.[5] This is not a useful distribution with which to work.

This sort of situation is not unfamiliar. In his Lectures on Gas Theory [1], Boltzmann considers the case of a homogeneous gas in a vessel. He defines a function $f(\vec{v}, t)$ – the so called 'density of velocities' – by requiring $f(\vec{v}, t)|d\vec{v}|$ to be the number of particles in a (fixed) region of space whose velocities lie in the region $d\vec{v}$ at time t. Let us suppose we are dealing with point particles, and let us suppose that the region $d\vec{v}$ is a sphere centered around a fixed velocity $\vec{v}$. Suppose N is the number of particles in the region of space in question with velocity $\vec{v}$ at time t. Then we must have:

$$\lim_{|d\vec{v}|\to 0} f(\vec{v}, t)|d\vec{v}| = N.$$

If $N \neq 0$, no possible value of $f(\vec{v}, t)$ can make this equation true, and if $N = 0$, every value of $f(\vec{v}, t)$ makes this equation true. So our definition has broken down.

Boltzmann's solution to think of $d\vec{v}$ as a *fixed*, tiny cell of velocity space centered around $\vec{v}$. Because $d\vec{v}$ is fixed, it is not the sort of thing we can send to 0. Consequently, the equation

$$f(\vec{v}, t)|d\vec{v}| = N$$

defines a unique value of $f(\vec{v}, t)$. Boltzmann's idea is that we should choose our cells $d\vec{v}$ in such a way that the resulting distribution of velocities $f(\vec{v}, t)$ is more or less continuous – so that when we move from a cell of velocity space to an 'adjacent' cell, $f(\vec{v}, t)$ does not change too much. That it is possible to choose cells $d\vec{v}$ in this way is a substantive assumption Boltzmann is happy to make. So long as this assumption is made, the quantity $f(\vec{v}, t)$ is well-defined enough for Boltzmann to work with.

We mimic this move. We divide phase space into small cells. Let dX be the cell containing X. We then define

$$\psi(X) = \frac{\tau(dX)}{T|dX|}. \tag{2}$$

The idea here is to choose the cells in such a way that ψ is more or less continuous – so that when we move from one of these cells to an adjacent one, $\psi(X)$ does not change too much. That it is possible to do this is a substantive assumption that

[5] In a finite amount of time, the path our system traces through phase space will generally consist of finitely many continuous curves, each of finite length. It may be shown to follow from this that the set of points through which the system does not pass is an open set, possibly plus some finite set of points, and that the set of points through which the system does pass is of measure 0. If we take a point X through which the system does not pass, it follows that, with finitely many exceptions, there will be some open ball centered around that point through which the system does not pass. And so when dX is sufficiently small, we will have $\tau(dX) = 0$. Thus, for sufficiently small dX, $\tau(dX)/(T|dX|) = 0$, and so it follows from our definition that $\psi(X) = 0$.

we simply take for granted in what follows. So long as this assumption is made, the quantity $\psi(X)$ is well-defined enough to work with.

With all this in mind, we formulate the following:

> **Main Definition I** (Tentative): For a system to be in a state described by the *canonical distribution* is for it to be the case that, for sufficiently large T, the distribution ψ defined by (2) is approximately equal[6] to the canonical distribution.

(We leave open the possibility that different coarse grainings may be required at different times.) It is only reasonable to talk about approximate equality in our definition for two reasons – first, the function ψ is coarse grained, and so cannot be identical with the canonical distribution, and second, there is no guarantee that in any finite time the system will have explored all the regions of phase space in exactly the correct proportions.[7]

A natural generalization of the previous definition is possible:

> **Main Definition II** (Tenative): For a system to be in a state described by the distribution ψ is for it to be the case that, for sufficiently large T, the distribution ψ defined by (2) is approximately equal to ψ.

There are several reasons why I am only prepared to call such definitions 'tentative'. I shall outline one. There are a number of ways we can think of a heat bath. According to one point of view, a heat bath is in a determinate microstate at any point in time, and so at the moment our system is placed in contact with a heat bath there are a variety of microstates in which the heat bath could be. Consequently, there will be different distributions ψ depending on the particular initial state of the heat bath. So rather than making a claim about a single ψ, our definitions should make a claim about the *most probable* distribution ψ (in some appropriately defined sense.) Whether this modification is necessary is not clear to me. I will not try to decide this issue here, but will simply call the definitions 'tentative', to acknowledge that their details may need to be altered to accommodate different ways of thinking about a heat bath.

Let us put all this to the side for now, and make some general remarks. In these definitions, our ensemble is the set microstates through which the system would pass were it left in contact with the reservoir indefinitely. To say that a system can be described by the canonical distribution is then to make a statistical claim about the (probable) structure of this time-ensemble. Note that one advantage of working with these definitions is that the focus from the start is on time averages. If time-averages are so important – as the huge amount of effort directed towards understanding the relationship between time and ensemble averages suggests –

[6]For our purposes, it suffices to say that two functions f_1 and f_2 on phase space are approximately equal iff the phase space integral $\int dX|f_1(X) - f_2(X)|$ is very small.

[7]One might wonder whether the distributions defined in (2) converge (in some sense) to the canonical distribution as $T \to \infty$. A requirement like this is possible, but because it does not add anything to the present discussion, I omit it for now.

it seems much more efficient to just *define* the canonical ensemble as a time-ensemble of sorts. This is not to suggest that by redefining words we can avoid any of the major philosophical problems with equilibrium statistical mechanics. All the old problems still exist, though many of them will need to be formulated in different language. The hope, however, is that better definitions can give us a better perspective on what these problems are, and how deep they run.

To see an example of an old problem in a new form, note that to describe a system with the canonical distribution (as defined in Main Definition I) is now to make a substantive assertion about the system. It is not a-priori obvious that it is ever correct to describe any system with the canonical distribution, let alone the multitude of systems to which the canonical distribution is actually applied. This is a very serious issue to which we shall return shortly.

One final remark: according to our definition, to describe a system with the canonical distribution is to make a claim about the (very) long term behavior of the system. As such, the values of observables calculated using the canonical distribution are to be thought of as time-averaged expectation values. The time-averaged expectation value of an observable can be quite different from the value of the observable at any moment of time. However, supposing we can show that the variance of the observable over time is small, we can at least assert that the observed value will *probably* be approximately equal to the time-averaged expectation value.[8] The observed value probably being approximately equal to the time-averaged expectation value is compatible with the observed value being radically different from the time-averaged expectation value. But so long as we realize that the canonical distribution only gives us the probable value of an observable[9], there is no real problem with the fact that the actual value of the observable might be very different, unless we insist on demanding from the canonical ensemble something that it is simply not designed to give.

3 Heat Baths and Canonical States.

In the previous section, we gave a tentative definition for what it means for a system to be in a state described by the canonical distribution. Our next question is whether systems in equilibrium with heat baths really do have states described by the canonical distribution. Does the canonical distribution accurately describe the set of microstates through which a system passes when it is in contact with a heat bath?

This is a tremendously difficult question. Some of these difficulties are of a mathematical sort, but others are more conceptual. For instance, we are interested

[8] An analagous move is made in [9], albeit with respect to ensemble averages and the microcanonical distribution.

[9] Here, we are talking about probability derived from the time-ensemble in the way described. There may be other notions of probability at play, due to the fact that the system in question will be a part of other ensembles or collectives. The canonical ensemble makes no claim to tell us anything about those other ensembles.

in studying a system interacting with a heat bath – but precisely what sort of system is a heat bath? A heat bath must contain an infinite amount of energy if it is to be possible that it produce arbitrarily large energy fluctuations in the system with which it interacts, as the canonical distribution predicts. In addition, a heat bath must be able to absorb or give up an arbitrarily large amount of heat without changing its own temperature. A heat bath is therefore an idealization. The question, then, is not what the microscopic structure of actual heat baths are, for there are no such things. Instead, the question is what sort of model we should use for heat baths.

The basic idea with which most authors begin is that a heat bath should be a system containing a countable infinity of particles spread out over a well defined, infinite region of space (where these conditions are, of course, not intended to be sufficient.) For instance, one might take a heat bath to be an infinitely long chain of harmonic oscillators. Provided the initial conditions of these particles are chosen correctly, such a system will contain an infinite amount of energy. Because heat capacity is an extensive quantity, the heat capacity of such a system will also be infinite. Thus, we have what is at least a prima facie candidate for a heat bath.

Let us be even more specific. Take an infinitely long chain of harmonic oscillators, ordered like the integers. Take some finite subset of adjacent oscillators (e.g., the 1st, 2nd, .., nth.) This finite sub-chain of oscillators will be our *system*, and the remainder of the oscillators will be our *heat bath*. Does this system actually behave in the way that we would expect a system in a heat bath to behave? This mathematical question has been studied in great detail. One of the pioneering works is Ford, Kac, and Mazur [5]; subsequent important contributions include Heurta and Robertson [6] [7], Davies [3], and Tegmark and Yeh [11].

Unfortunately, the upshot of this body of literature is difficult to assess. First, on the terms on which much of this literature proceeds, the theorems proven there show that our candidate for a heat bath does *not* behave in the desired way, at least in general. But second, it is not clear that the theorems proved in these papers are the sorts of theorems one would actually want to prove if one's goal was to show that the heat-bath-candidate behaves in the desired way. I will address each of these points in turn. My conclusion is that the question of how to best model a heat bath is unresolved. Because of this, the question of whether a system in equilibrium with a heat bath should be described by the canonical distribution (invoking Main Definition I) remains wide open.

First, let us consider the results of this literature on its own terms. In this literature, it is assumed that we can talk about a probability distribution that describes the system *at any given time*, and that this distribution undergoes the usual time evolution determined by the Hamiltonian. This way of thinking about what it is to describe a system with a probability distribution goes against our Main Definition II, according to which the probability distribution describes a time ensemble rather than an ensemble at an instant[10], but it is in this sense that

[10]The fact that we may want to describe the heat bath with a probability distribution at a given instant does not contradict the idea that the canonical distribution first and foremost makes a claim

we consider the literature on its own terms. In [6], Huerta and Robertson study a system with the Hamiltonian

$$H = \sum_{n=-\infty}^{n=+\infty} [\frac{p_n^2}{2m} + \frac{k(x_{n+1} - x_n)^2}{2}]$$

They assume that the initial positions and momenta of particles both in the system and in the bath are given by uncorrelated Guassians. In particular, they write:

$$\rho(\{x\}, \{p\}, t = 0) = \prod_{n \in \{1 \ldots N\}} \frac{\{\exp -[x_n(0) - u_n]^2/2\alpha^2\}}{\alpha(2\pi)^{1/2}} \cdot \frac{\exp\{-[p_n(0) - v_n]^2/2\delta^2\}}{\delta(2\pi)^{1/2}}$$
$$\times \prod_{n \notin \{1 \ldots N\}} \frac{\exp\{-x_n(0)^2/2\epsilon^2\}}{\epsilon(2\pi)^{1/2}} \cdot \frac{\exp\{-p_n(0)^2/2\zeta^2\}}{\zeta(2\pi)^{1/2}}$$

where it is assumed that $\alpha, \delta, \epsilon, \zeta$ are all non-zero. Note that the above formula allows the initial conditions of the particles in the system to be skewed. They then show that

$$\lim_{t \to \infty} \langle p_r(t) p_{r+1}(t) \rangle \neq 0$$

(where $\langle \cdot \rangle$ denotes the phase space average with respect to the measure ρ at time t.) Because there are no correlations between the momenta of different particles in the canonical distribution, it follows that the system does *not* relax into the canonical distribution.[11] Thus, we cannot think of an infinite chain of oscillators with initial conditions given by uncorrelated Gaussians as an effective model of a heat bath.[12]

The argument of Tegmark and Yeh [11] is a little more general, but broadly similar conclusions are reached. In the conclusion of their paper, when they discuss distributions proportional to $e^{-H/kT}$ (such as the canonical distribution), they claim that '*no completely solvable heat bath model has ever been found that explicitly evolves multiparticle* ($n > 1$) *systems into such states*' (p. 359).

There is (almost) no questioning the mathematics behind the calculations of Huerta and Robertson or Tegmark and Yeh. What is less clear is what these calculations actually show. Both pairs of authors assume that what needs to be determined is whether the probability distribution associated with the initial state of a particular open system evolves into the canonical distribution as $t \to \infty$. Indeed, this is the approach of a large body of literature to which these authors' papers belong.

But it is not clear what is demonstrated by the fact that an initial probability distribution does (or does not) evolve into a particular final probability distribution as $t \to \infty$. For according to Main Definition II, to describe a system using

about time ensembles.

[11] The argument in §6 of [6] applies only to particles in the bath (a restriction imposed 'for simplicity'), but it is easy to generalize this argument to cover particles in the system.

[12] Huerta and Robertson do show that *some* features of equilibrium are obtained in the long run, such as equipartition of energy.

a probability distribution ρ is to make a claim about how the system evolves as $t \to \infty$. Given these definitions, it does not make sense to interpret the relaxation of a system into equilibrium as the evolution of one probability distribution to another.[13] Indeed, it does not make sense to talk about a system in contact with a heat bath changing its state ρ at all. Insofar as our goal is to intepret and justify the machinery of *equilibrium* statistical mechanics, this is not worrisome.

The net result of all this is that is difficult to interpret the calculations of Huerta et. al as telling us anything meaningful about the set of microstates through which a system passes when it remains in contact with a putative heat-bath. Because of this, it is difficult to draw any conclusions from this literature about what does or does not count as a good model of a heat bath. Any serious attack of this question needs to examine the way in which *particular* initial systems (rather than ensembles of initial systems) move through phase space when placed in contact with a putative heat bath (or ensemble of putative heat baths.) To my knowledge such matters remain unexplored, and we do not explore them here. The important point is that the manner in which one approaches the *physics* question of how to model a heat bath is determined in part by the manner in which one approaches the *philosophical* question of what kind of thing the canonical distribution is.

4 Concluding Remarks.

The question of why we are right to describe systems in equilibrium with heat baths using the canonical distribution remains open. Even once this problem is addressed, other difficult issues then require attention. For instance – why it is possible to treat a physical system in equilibrium with a large, finite reservoir of heat as if it were in contact with an *infinite* reservoir of heat? And why it is permissible to use the canonical distribution to describe systems that are not in contact with any sort of reservoir of heat, as is often done? (The fact that the microcanonical distribution and canonical distribution approach each other in the thermodynamic limit is often claimed to justify the use of the canonical distribution for closed systems, but it is far from clear that this argument is convincing.) Regardless of all this, one moral that I hope can be drawn from this paper is that thinking about what sort of thing the canonical distribution is can help to clarify the mathematical questions that must be addressed in order to then justify its use. This is at least a modest sort of progress.

[13] Note, however, that it does make sense to talk about a change in ρ when a heat bath is replaced with a different bath, insofar as one can talk about the states through which the system would have passed were it to have remained in contact with the original heat bath, as well as the states through which it will pass now that it is in contact with a new heat bath. Thus, there is the possibility of talking about the way in which the state changes during certain quasi-static processes.

References

[1] Boltzmann, L., *Lectures on Gas Theory*, Dover Publications, 1964.

[2] Callendar, C., 'Reducing Thermodynamics to Statistical Mechanics: The Case of Entropy', *The Journal of Philosophy*, Vol. 96, No. 7 (July 1999), 348-373.

[3] Davies, E. B., 'The Harmonic Oscillator in a Heat Bath', *Commun. Math. Phys.*, Vol 33 (1973), 171-186.

[4] Earman, J., and Redei, M., 'Why Ergodic Theory Does Not Explain the Success of Statistical Mechanics', *British Journal for the Philosophy of Science*, Vol. 47, (1996), 63-78.

[5] Ford, G., Kac, M., and Mazur, P., 'Statistical Mechanics of Assemblies of Coupled Oscillators', *Journal of Mathematical Physics*, Vol 6, No. 4 (April 1965), 504-515.

[6] Huerta, M and Robertson, H., 'Entropy, Information Theory and the Approach to Equilibrium of Coupled Harmonic Oscillator Systems', *Journal of Statistical Physics*, Vol 1, No. 3 (1969), 393-414.

[7] Huerta, M and Robertson, H., 'Approach to Equilibrium of Coupled Harmonic Oscillator Systems II', *Journal of Statistical Physics*, Vol 3, No. 2 (1971), 171-189.

[8] Gibbs, J. Willard., *Elementary Principles in Statistical Mechanics*, Ox Bow Press, 1981.

[9] Malament, D.,and Zabell, S., 'Why Gibbs Phase Averages Work – The Role of Ergodic Theory', *Philosophy of Science*, Vol 47, 339-349.

[10] Sklar, L., *Physics and Chance*, Cambridge University Press, 1993.

[11] Tegmark, M. and Yeh, L., 'Steady States of Harmonic Oscillator Chains and Shortcomings of Harmonic Heat Baths', *Physica A*, 202 (1994), 342-262.