International Journal of Theoretical Physics, Vol. 18, No. 5, 1979

On the Meaning of the Canonical Ensemble

Rafael Sorkin[1]

Department of Applied Mathematics and Astronomy, University College, Cardiff CF1 1XL, United Kingdom

Received March 15, 1979

Thermal equilibrium between (quantum) systems is taken to mean stability for the combined system. Necessary and sufficient conditions for such stability are found and used to show that any system in equilibrium with suitably complex second system ("heat bath") will be characterized by a canonical ensemble. Thus the notion of temperature is derived directly from that of equilibrium, without, for example, recourse to microcanonical ensembles or information theory. Discussed briefly are the generalization of these results to grand canonical ensembles and their application to the equilibrium between a black hole and the surrounding radiation field.

Although in most cases the canonical ensemble adequately describes thermodynamic equilibrium, there are important physical systems for which it cannot be used. For example, the defining relations

$$\rho = e^{-\beta H}/Z$$
$$Z = \mathrm{tr}\, e^{-\beta H} \tag{1}$$

imply on the one hand

$$\begin{aligned}(\log Z)'' &= Z''/Z - (Z'/Z)^2 \\ &= \langle H^2\rangle - \langle H\rangle^2 \\ &= \Delta H^2\end{aligned}$$

and on the other

$$(\log Z)'' = -\frac{d}{d\beta}\langle H\rangle = T^2\frac{d\langle H\rangle}{dT}$$

[1]Present address: Enrico Fermi Institute, University of Chicago, Chicago, Illinois 60637.

0020-7748/79/0500-0309$03.00/0

whence the heat capacity $d\langle H\rangle/dT$ would always be $\geqslant 0$. But in fact gravitationally bound systems such as stars and black holes are known to display *negative* heat capacities. Does this mean that the canonical ensemble is merely a mathematically convenient form, sometimes useful but sometimes to be replaced by other constructions such as the microcanonical ensemble? Or is it the only true representative of thermal equilibrium so that systems like stars cannot be considered as truly thermodynamic?

The present paper will approach these questions neither from the information-theoretic viewpoint, nor by means of some infinite replication of the system concerned. Rather (cf. Born, 1964) we will adopt as basic the idea of equilibrium between a single pair of *individual* (*quantum mechanical*) systems, a notion which itself is not obviously free from ambiguity as appears, e.g., in discussions of thermal equilibrium between relatively moving bodies.

That two systems are in thermal equilibrium means, more or less by definition, that their states do not change when heat is allowed to flow between them; and of course it should not matter exactly by what means the systems are thermally coupled. Aside from equating the idea of "allowing heat to flow" to that of introducing an arbitrary weak coupling between the two systems, we will adopt the requirement just stated as the criterion of thermal equilibrium. It will follow then first of all that only certain states of a given system (including the canonical and microcanonical ensembles) are capable of being in equilibrium at all, no matter with what other system, and secondly (though less rigorously) that if a system is sufficiently complex to be called a "heat bath" then *only* canonical states can be in equilibrium with it.

As just proposed, our notion of thermal equilibrium between individual systems reduces in effect to a kind of stability for the combined system. In order to formulate this precisely let us introduce, for a given system S, the corresponding Hilbert space $\mathfrak{h}$, (unperturbed) Hamiltonian H, and space P of statistical states (density matrices). Recall also that, because of Liouville's theorem, or rather the unitarity of the time-evolution operator, we cannot, without some sort of "coarse graining" ask that a finite system actually approach equilibrium as $t\to\infty$. Instead, we will take "stability" to mean only that a small change in H produces a change in ρ which remains small for all time.[2]

[2]This might seem to be a needlessly weak requirement for thermal stability. Thus, because a pair of systems at the same temperature should, when thermally coupled, settle down into a nearby equilibrium state of the combined system, we certainly could ask such a nearby state *exist*, even if (because of the course-graining problem) we cannot see the settling down mathematically. However, it turns out—at least when spec(H) is discrete—that Definition 1 will already entail the existence of nearby stationary states of the perturbed system.

Definition 1. $\rho \in P$ is an *h-stable* equilibrium state of S iff for every neighborhood N of ρ there are neighborhoods M of ρ and W of H such that $\forall \hat{H} \in W$ the evolution generated by $\hat{H}$ carries M into N for all time.[3] (In particular ρ itself must be stationary.)

Definition 2. Two systems S_A, S_B with Hamiltonians H_A, H_B and in states ρ_A, ρ_B are in *h equilibrium* if the combined state $\rho = \rho_A \otimes \rho_B$ is an h-stable equilibrium state of the combined system, S.

Mathematically each word "neighborhood" in Definition 1 presupposes a topology. When the space $\mathfrak{h}$ is finite dimensional, these topologies are unique (and obvious), but unfortunately many thermodynamic systems —not least among them heat baths—have unbounded Hamiltonians and (therefore) infinite-dimensional Hilbert spaces. For general $\mathfrak{h}$ we can introduce what seem to be convenient topologies as follows.

Let $\mathfrak{L}(\mathfrak{h})$ be the space of bounded linear operators in $\mathfrak{h}$. Having identified P with a subset of $\mathfrak{L}(\mathfrak{h})$ [namely, the set $P(\mathfrak{h}) := \{\rho \in \mathfrak{L}(\mathfrak{h}) | \rho \geqslant 0$ and $\operatorname{tr}\rho = 1\}$ of normalized density matrices] we can use for it the topology it thereby inherits.[4] The possible Hamiltonians, however, are drawn not only from $\mathfrak{L}(\mathfrak{h})$, but from the space $\mathfrak{L}^{USA}(\mathfrak{h})$ of not necessarily bounded self-joint operators on $\mathfrak{h}$. In this space, we can consider two operators to be close to each other when their difference is a bounded operator of small norm.[5]

Rewritten in accord with these choices our first definition would read as follows:

> $\rho \in P(\mathfrak{h})$ is an h-stable equilibrium state iff $\forall \epsilon > 0$, $\exists \delta_1, \delta_2 > 0$, $\forall \hat{P} \in P(\mathfrak{h})$, $\forall \hat{H} \in \mathfrak{L}^{USA}(\mathfrak{h})$, $\|\hat{\rho} - \rho\| < \delta_1$, and $\|\hat{H} - H\| < \delta_2 \Rightarrow$ $\forall t \|\hat{U}(t)\hat{\rho}\hat{U}(-t) - \rho\| < \epsilon$, where $\hat{U}(t) = \exp(-i\hat{H}t)$.

However, I claim that the following simpler version is equivalent to that just stated:

[3]Some people use the term "structural stability" for a concept of this type, which envisions perturbing the equations of motion as well as the initial conditions. For a quantum system (but not necessarily for its classical analog!), the former sort of perturbation is the *only* sort that is relevant since the unitarity of the time-evolution operator ensures that *any* stationary state is stable under perturbation of initial conditions alone.

[4]This turns out to be also the topology of pointwise convergence on elements of $\mathfrak{L}(\mathfrak{h})$ when $\rho \in P$ acts on $\mathfrak{L}(\mathfrak{h})$ by $\rho(A) = \operatorname{tr}(\rho A)$. Using this fact one can check that when $\mathfrak{h} = \mathfrak{h}_A \otimes \mathfrak{h}_B$ the map, tr_B, from $P(\mathfrak{h})$ to $P(\mathfrak{h}_A)$ gotten by "tracing out" the $\mathfrak{h}_B$ variables is continuous. In particular, this ensures that Definition 2 implies that not only ρ but also the relative state, $\operatorname{tr}_B\rho$, describing S_A is stable in the required sense. Conversely, any topology for P ought to have the property just described if it is to be used in the context of Definition 2.

[5]The "gap topology" of Kato (1966) might be more natural but would, I think, lead to the same results in any case.

Definition 1'. $\rho \in P(\mathfrak{h})$ is an h-stable equilibrium state iff $\forall \epsilon > 0$, $\exists \delta > 0$, $\forall \hat{H} \in \mathfrak{L}^{USA}(\mathfrak{h})$, $\|\hat{H} - H\| < \delta \Rightarrow \forall t \|\hat{U}(t)\rho\hat{U}(-t) - \rho\| < \epsilon$, where $\hat{U}(t) = \exp(-i\hat{H}t)$.

Proof. We must show the present condition implies the previous one (the opposite implication being obvious). So suppose ρ is h stable in the present sense. Then for all t,

$$\|\hat{U}(t)\hat{\rho}U(-t) - \rho\| \leqslant \|\hat{U}(t)(\hat{\rho} - \rho)\hat{U}(-t)\| + \|\hat{U}(t)\rho\hat{U}(-t) - \rho\|$$

$$= \|\hat{\rho} - \rho\| + \|\hat{U}\rho\hat{U}^{-1} - \rho\|$$

Given $\epsilon > 0$, we can by hypothesis find δ so that the second term is $< \epsilon/2$ for all $\hat{H}$ such that $\|\hat{H} - H\| < \delta$. Thus the previous condition is satisfied with $\delta_1 = \epsilon/2$ and $\delta_2 = \delta$. ■

Doubtless the present framework is neither mathematically nor physically the best possible. In the way of generalization one might want to replace $\mathfrak{L}(\mathfrak{h})$ by an arbitrary "factor" or even an arbitrary C^* algebra, $P(\mathfrak{h})$ by the set of all states on that algebra, etc. It would be interesting to see to what extent the results presented here carry over to such a more general context, which may well be needed, e.g., by quantum field theory.[6] In places we will make further restrictions for mathematical convenience, in particular that the spectrum of the Hamiltonian be discrete.

Theorem 1. If ρ is an h-stable equilibrium state then ρ is a function of the Hamiltonian H.

Proof. Since ρ is stationary $U(t)\rho U(-t) = \rho$ [which fact, namely, that $U(t)$ commutes with ρ, we will write as "$U(t) \natural \rho$"] for all t. Let $W \in \mathfrak{L}^{SA}(\mathfrak{h}) = \{T \in \mathfrak{L}(\mathfrak{h}) | T = T^*\}$ and suppose ϵ and δ are as in Definition 1'. Set $V = \frac{1}{2}\delta \|W\|^{-1} W$. Then $\|V\| < \delta$ so that by hypothesis $\forall t \|\hat{U}(t)\rho\hat{U}(-t) - \rho\| < \epsilon$, where $\hat{U}(t) = \exp(-i\hat{H}t)$ and $\hat{H} = H + V$. Suppose now that $W \natural H$. Then[7] also $V \natural H$ and $\hat{U}(t) = e^{-iVt}U(t)$ so that $\hat{U}(t)\rho\hat{U}(-t) = e^{-iVt}\rho e^{iVt}$. We have then $\forall \epsilon \exists \delta \forall t \|e^{-iVt}\rho e^{iVt} - \rho\| < \epsilon$, whence, since $V \propto W$, the same holds with W replacing V. But now the arbitrariness of ϵ shows $\forall t e^{-iWt} \natural \rho$, which is equivalent [see, e.g., Theorem VIII.13 of Read and Simon (1972)] to $W \natural \rho$. Noting that for any $T \in \mathfrak{L}(\mathfrak{h})$, $T \natural H$ iff both its real and imaginary

[6] As it happens, much work has been devoted recently to finding results in this direction (albeit in the context of the infinite thermodynamic limit) (Bratelli, 1978a; Bratelli et al., 1978b).

[7] See Riesz and Sz.-Nagy (1955) for this and for the definition of $A \natural B$ when A is unbounded.

parts $\natural H$ we conclude that $\forall T \in \mathfrak{L}(\mathfrak{h}) T \natural H \Rightarrow T \natural \rho$, which means[8] that ρ is a *function* of H. ∎

Remark 1. This proof did not use the specific topology of P in any important way. Moreover it would work as well using any other of the usual topologies for $\mathfrak{L}^{USA}(\mathfrak{h})$ since thay are all *weaker* than ours.

> *Theorem 2.* If f: $\mathbb{R} \to \mathbb{R}$ is continuous and $\rho = f(H)$ belongs to $P(\mathfrak{h})$, then ρ is an h-stable equilibrium state.

Proof. Since by definition $\mathrm{tr}\rho = 1 < \infty$, $f(x) \to 0$ as $|x| \to \infty$. [More accurately, since only $f \restriction \mathrm{spec}(H)$ affects $f(H)$, f can only be *taken* to have this property.] For such an f the map $A \to f(A)$ of $\mathfrak{L}^{USA}$ into $\mathfrak{L}^{SA}$ is continuous [this being Theorem VIII.20 of Reed and Simon (1972) since it follows easily from Theorems (2.14) and (2.20) of Chapter IV of Kato (1962) that any sequence H_n which approaches H in our sense also approaches H in the "norm-resolvent" sense.] In particular, given ϵ there is δ such that $\|\hat{H} - H\| < \delta$ implies $\|f(\hat{H}) - f(H)\| < \epsilon/2$. But $f(\hat{H})$, being a function of $\hat{H}$, is stationary, (though not necessarily of unit, or even of finite, trace) under the evolution $\hat{U}(t)$ generated by $\hat{H}$:

$$\hat{U}(t) \natural f(\hat{H})$$

so that for all t [and writing $\hat{\rho}(t) = \hat{U}(t) \rho \hat{U}(-t)$]

$$\begin{aligned} \|\hat{\rho}(t) - f(\hat{H})\| &= \|\hat{U}(t)[\rho - f(\hat{H})]\hat{U}(-t)\| \\ &= \|\rho - f(\hat{H})\| \qquad \text{(since } \hat{U} \text{ is unitary)} \\ &= \|f(H) - f(\hat{H})\| \end{aligned}$$

Thus for all t,

$$\|\hat{\rho}(t) - \rho\| \leqslant \|\hat{\rho}(t) - f(\hat{H})\| + \|f(\hat{H}) - \rho\| = 2\|f(\hat{H}) - f(H)\| < \epsilon \qquad ∎$$

Corollary. If $\mathrm{spec}(H)$ is discrete then $\rho \in P$ is h stable iff $\rho = f(H)$ for some f: $\mathbb{R} \to \mathbb{R}$

Proof. If f: $\mathbb{R} \to \mathbb{R}$ and $\rho = f(H)$ then, without changing $f(H)$ we can redefine f freely on $\mathbb{R} \setminus \mathrm{spec}(H)$. Since $\mathrm{spec}(H)$ is discrete this can be done so as to make f continuous. ∎

[8] It may be that this follows only when $\mathfrak{h}$ is a separable Hilbert space; see Section 129 of Riesz and Sz.-Nagy (1955), which also proves it in this case.

Remark 2. With somewhat more effort we could prove the more general result that, without any restriction on the spectrum of H, ρ is h stable$\Leftrightarrow\rho=f(H)$ for some *continuous* function f.

> *Theorem 3.* Let S_A, S_B be two systems with Hamiltonians H_A, H_B and in states ρ_A, ρ_B. Suppose that H_A and H_B have discrete spectra. Then S_A and S_B are in h equilibrium if and only if (i) $\rho_A=f_A(H_A), \rho_B=f_B(H_B)$ for real-valued functions $f_{A,B}$, and (ii) $\forall\epsilon'_A$, $\epsilon''_A\in\text{spec}(H_A)$, $\forall\epsilon'_B$, $\epsilon''_B\in\text{spec}(H_B)$
>
> $$\epsilon''_A-\epsilon'_A=\epsilon''_B-\epsilon'_B\Rightarrow f_A(\epsilon''_A)/f_A(\epsilon'_A)=f_B(\epsilon''_B)/f_B(\epsilon'_B)$$
>
> (the latter condition being understood in the obvious way when either of the denominators vanishes.) In particular, if ρ_A and ρ_B are canonical states at the same temperature [see equation (1)] then S_A and S_B are in h equilibrium.

Proof. Let S be the combined system with Hilbert space $\mathfrak{h}=\mathfrak{h}_A\otimes\mathfrak{h}_B$ and Hamiltonian $H=H_A+H_B$ (or more properly, $H_A\otimes 1+1\otimes H_B$). By definition S_A and S_B are in h equilibrium iff $\rho=\rho_A\otimes\rho_B$ is h stable. In particular this means that ρ_A and ρ_B are separately h stable (just take V of the form $V_A\otimes 1$ [resp. $1\otimes V_B$] in Definition 1′] whence according to Theorem 1 $\rho_A=f_A(H_A), \rho_B=f_B(H_B)$. Assuming this, h stability of ρ amounts to the condition

$$f_A(H_A)\otimes f_B(H_B)=f(H_A+H_B) \tag{2}$$

To see what this means write H_A in the form

$$H_A=\sum_l \lambda_l E_l$$

where $\{\lambda_l\}$ are the eigenvalues of H_A and $\{E_l\}$ the projections onto the corresponding eigensubspaces [this is possible by our assumption on spec(H_A)]. Similary, write

$$H_B=\sum_m \mu_m F_m$$

Putting $E_{lm}=E_l\otimes F_m$ we get, because $\Sigma E_l=1$, $\Sigma F_m=1$,

$$\begin{aligned} H&=\sum_l \lambda_l E_l\otimes 1+\sum_m \mu_m 1\otimes F_m=\sum_{lm}\lambda_l E_l\otimes F_m+\sum_{ml}\mu_m E_l\otimes F_m \\ &=\sum_{lm}(\lambda_l+\mu_m)E_{lm}=\sum_n \nu_n G_n \end{aligned} \tag{3}$$

where the ν_n range over the distinct values $\lambda_l + \mu_m$ and where

$$G_n = \sum_{\lambda_l + \mu_m = \nu_n} E_{lm} \tag{4}$$

Furthermore, $E_{lm}E_{l'm'} = E_l E_{l'} \otimes F_m F_{m'} = \delta_{ll'}\delta_{mm'}E_{lm}$, $E_{lm}{}^* = E_{lm}$, and

$$\sum_{lm} E_{lm} = \left(\sum_l E_l\right) \otimes \left(\sum_m F_m\right) = 1$$

so that the G_n are a resolution of unity and (3) is in fact the spectral decomposition of H. Now $\rho_A = f_A(H_A) = \Sigma f_A(\lambda_l)E_l$ (and similarly ρ_B) so that

$$\rho = \rho_A \otimes \rho_B = \sum_{lm} f_A(\lambda_l) f_B(\mu_m) E_{lm}$$

Comparing with (3) and (4) we see that ρ is of the form $f(H)$, namely, of the form

$$\rho = \Sigma f(\nu_n) G_n$$

if and only if one can choose f so that

$$\forall n \forall l \forall m \; \lambda_l + \mu_m = \nu_n \Rightarrow f_A(\lambda_l) f_B(\mu_m) = f(\nu_n).$$

Clearly (ii) is the necessary and sufficient condition that this be possible.

Finally, if ρ_A and ρ_B are canonical with positive temperature β^{-1}, then

$$f_A(\epsilon_A'')/f_A(\epsilon_A') = \epsilon^{-\beta(\epsilon_A'' - \epsilon_A')}$$

which equals $f_B(\epsilon_B'')/f_B(\epsilon_B')$ when $\epsilon_A'' - \epsilon_A' = \epsilon_B'' - \epsilon_B'$. (In case $T = 0$ the canonical state ρ is just $G_0/\mathrm{tr}G_0$, where ν_0 is the least eigenvalue of H, and clearly is the product of the $T = 0$ canonical states of H_A and H_B.) ■

Remark 3. It seems likely that, at the cost of some measure-theoretic complication, one could establish Theorem 3 in full generality except that, following Remark 2, f_A and f_B would be required to be continuous in condition (i).

Theorem 3 shows that the canonical ensemble (when it exists) is a sort of universal equilibrium state. On the other hand, it is also clear from Theorem 3 that for most pairs of simple systems S_A, S_B, there will be states ρ_A, ρ_B of mutual h equilibrium neither of which is canonical, just as one would expect on physical grounds. By the same token, however, one might

also expect that a system in contact with a "heat bath" would be characterized by a definite temperature, and therefore possibly by a canonical state at that temperature.

At this point I would like to define a heat bath as a system whose Hamiltonian's spectrum is of the form $[a, \infty)$,[9] but unfortunately I do not know how to make sense of such a case. [Perhaps by replacing $\mathfrak{L}(\mathfrak{h})$ by a type II_1 factor?] Let us therefore work with the less rigorous notion of a Hamiltonian whose spectrum is defined by a *level-density* function μ and assume that support $(\mu)=[a, \infty)$. We are then trying to describe a heat bath as a system with a density of energy levels which is very high and becomes infinite in, say, the limit of infinite volume. (In practice, the criterion would be that the level spacing of H_B be much smaller than that of H_A.) Notice that not only systems one would usually think of as heat baths have this character but also much simpler systems such as a single free particle in a very large box.

> *"Theorem" 4.* If a heat bath S_B is in h equilibrium with a second system S_A for which $\mathrm{spec}(H_A)$ is discrete, then ρ_A is a canonical ensemble.

Remark 4. The term "heat bath" might carry the misleading connotation of "heat bath at temperature T," but as used here it has to do with the *construction* of S_B alone, not with its state.

"Proof." From Theorem 3 we can assume $\rho_A = f_A(H_A), \rho_B = f_B(H_B)$. Assume also that "$\mathrm{spec}(H_B)=[a, \infty)$" as discussed above. Notice that $f_B(\epsilon)$ is the occupation probability per state at energy ϵ, so that, e.g., $\mathrm{tr}\rho_B = \int_a^\infty f_B(x)\mu_B(x)dx$. We could let f_B be any measurable (with respect to the measure $\int \mu_B(x)dx$] function and convert to continuous functions by convolution with smooth functions of compact support, but in accord with the general level of rigor, and with Remark 3, let us simply take f_B to be continuous. As for ρ_A we can number the eigenvalues of H_A in increasing order, $\epsilon_0, \epsilon_1, \epsilon_2, \ldots$, and write p_k for $f_A(\epsilon_k)$. Condition (ii) of Theorem 3 becomes then (dropping the subscript on 'f_B')

$$\forall j,k;\ \forall x,y \geqslant a, \qquad \epsilon_k - \epsilon_j = y - x$$
$$\Rightarrow p_k f(x) = p_j f(y) \tag{5}$$

I claim that if $p_j = 0$ for any j then all subsequent p's vanish. In fact if $k \geqslant j$ and if $\Delta\epsilon$: $= \epsilon_k - \epsilon_j$, then taking $y = x + \Delta\epsilon$ in (5),

$$p_k f(x) = p_j f(y) = 0 \qquad \forall x \geqslant a$$

[9] Or $(-\infty, a]$ in the case of negative temperatures.

Since $f \not\equiv 0$ this implies that $p_k = 0$ as claimed. In particular, we see that $p_0 > 0$. It *can* happen that $p_1 = 0$, but in that case all other p's vanish as well so that ρ_A represents a canonical ensemble with $T = 0$, and we are done.

Suppose then that $p_1 > 0$ and set $\Delta\epsilon = \epsilon_1 - \epsilon_0, \beta = -\log(p_1/p_0)/\Delta\epsilon$, and $f(x) = g(x)e^{-\beta x}$. Equation (5) (with $j = 0, k = 1$, and $y = x + \Delta\epsilon$) becomes the statement that g is periodic with period $\Delta\epsilon$:

$$\forall x \geqslant 0 \qquad g(x + \Delta\epsilon) = g(x) \tag{6}$$

Furthermore, if any $p_k = 0 (\Rightarrow k > 0)$, then again from (5) (with $j = 0$ and $y = x + \epsilon_k - \epsilon_0$)

$$f(y) = 0 \qquad \forall y \geqslant a + \epsilon_k - \epsilon_0$$

whence, because of the periodicity (6), $f \equiv 0$, which is impossible. Therefore $p_k > 0$ for all k, which allows us to repeat the derivation of (6) for all pairs j, k, getting

$$\forall j, k \; f(x) = g_{jk}(x)e^{-\beta_{jk}x} \tag{7}$$

where $\beta_{jk} = -\log(p_k/p_j)/(\epsilon_k - \epsilon_j)$ and g_{jk} is periodic with period $\epsilon_k - \epsilon_j$. Since f (and therefore g) is continuous the different expressions (7) will be compatible only if all the β's are equal, which in turn says that ρ_A is canonical. ■

If we apply this result to a heat bath in the ordinary sense, namely, an infinite collection of weakly coupled subsystems, all in thermal equilibrium with each other, then we conclude at once that each finite subsystem must be in a canonical state. Moreover, we can see to some extent why the usual approach based on a microcanonical ensemble for such a bath as a whole leads to this same conclusion. For according to Theorem 2, the microcanonical ensemble (more accurately any $\rho = f(H)$ where f is sharply peaked but continuous) is an h-stable state of the whole bath. *To the extent* that the subsystems behave independently, this *means*, by the terms of Definition 2, that each subsystem is in h equilibrium with the rest of the bath, and therefore in a canonical state as we have just remarked.

The question of independence just alluded to points to a possible loophole in the reasoning which culminated in Theorem 4. Even if, in Definition 2, $\rho = \rho_A \otimes \rho_B$ were not itself stable, ρ_A and ρ_B still might be deemed to be in thermal equilibrium as long as the *relative* states $\mathrm{tr}_B\rho$ and $\mathrm{tr}_A\rho$ remained stable. In other words, one would allow the coupling V to introduce essential correlations between S_A and S_B as long as these were undetectable by observing S_A and S_B separately. Luckily it turns out that such a weakening of Definition 3 would not of itself actually widen the class of h-equilibrium states. Perhaps, though, it might in the context of the following generalization, which in any case is much more important physically.

To couple S_A with S_B we have allowed ourselves any interaction whatsoever, whereas basic principles such as locality and conservation laws might in reality restrict our choice of V's. Suppose then that the possible interactions, V, are comprised in a (von Neumann) algebra, $\mathcal{O} \subset \mathcal{L}(\mathfrak{h})$. In the proof of Theorem 1 we get now

$$(V \in \mathcal{O} \text{ and } V \natural H) \Rightarrow V \natural \rho$$

By some early theorems of Dixmier (1969) this is the same as

$$V \natural \mathcal{O}' \cup \{H\} \Rightarrow V \natural \rho$$

(where $\mathcal{O}' := \{T \in \mathcal{L}(\mathfrak{h}) | \forall A \in \mathcal{O}, T \natural A\}$), i.e.,

$$\rho \natural (\mathcal{O}' \cup \{H\})'$$

$$\rho \in (\mathcal{O}' \cup \{H\})''$$

In other words ρ no longer need be "made from" H alone but from H together with those operators conserved by all possible interactions V. In the particularly simple case where these conserved operators (the elements of $\mathcal{O}'$) commute among themselves (and with H) and are additive for composite systems, we can expect to recover a suitable generalization of Theorem 4.

For example, if J is a single absolutely conserved additive quantity and if $\mathcal{O} = \{J\}'$ then we will have in equilibrium $J = J_A \otimes 1 + 1 \otimes J_B$, $\rho_A = f_A(H_a, J_A), \rho_B = f_B(H_B, J_B)$. The "joint spectrum" σ of H and J will be a subset of the H-J plane and (writing "ξ" for a point of this plane) the analog of condition (ii) of Theorem 3 will be

$$\xi'_A, \xi''_A \in \sigma_A, \xi'_B, \xi''_B \in \sigma_B \text{ and } \xi''_A - \xi'_A = \xi''_B - \xi'_B$$
$$\Rightarrow f_A(\xi''_A)/f_A(\xi'_A) = f_B(\xi''_B)/f_B(\xi'_B)$$

Again the "generalized canonical ensemble"

$$\rho = Z^{-1} e^{-\beta H - \alpha J}$$

is a universal solution and in many circumstances will be essentially the unique solution for ρ_A if ρ_B is suitably complex.

Finally let us return to one of our original questions and ask what becomes of Theorem 4 when S_A is, say, a black hole and S_B is the

surrounding radiation field, the combined system $S = S_A + S_B$ being enclosed in a box of volume V.[10]

For the black hole,

$$S(=\text{entropy}) \sim \text{area} \sim M^2$$

$$\beta \equiv T^{-1} = \partial S / \partial M \sim M, \qquad \frac{\partial^2 S}{\partial M^2} \sim 1 \tag{8}$$

and (taking $k \equiv 1$)

$$N(=\text{number of levels}) \sim e^S$$

so that

$$\text{level density} = \frac{\partial N}{\partial M} \sim \beta e^S \sim M e^{M^2}$$

while for the radiation

$$S \sim T^3 V, \qquad U \sim T^4 V$$

$$\Rightarrow S \sim (U^3 V)^{1/4}, \qquad \frac{\partial S}{\partial U} = \beta \sim U^{-1/4} V^{1/4} \tag{9}$$

$$\partial^2 S / \partial U^2 \sim -U^{-5/4} V^{1/4} \sim \beta U^{-1}$$

and as always

$$\frac{\partial N}{\partial U} \sim \beta e^S$$

For equilibrium $\beta_A = \beta_B = \beta$, i.e., $M \sim \beta$.

In the context of Theorem 4, the condition that S_B act as a heat bath for S_A is

$$\frac{\partial N_A}{\partial M} \ll \frac{\partial N_B}{\partial U} \tag{10}$$

$$\Leftrightarrow \beta e^{S_A} \ll \beta e^{S_B}$$

$$\Leftrightarrow S_A \ll S_B$$

$$\Leftrightarrow M^2 \ll T^3 V \sim U/T \sim UM$$

$$\Leftrightarrow M \ll U \tag{11}$$

[10] To avoid worrying about things like stimulated emission, one might want to surround the hole with a smaller box and think of *that* as S_A. This would hardly affect the analysis in the text.

By comparison the condition for thermodynamically stable equilibrium is that entropy be maximized with respect to energy exchange:

$$\frac{\partial^2 S_A}{\partial M^2} + \frac{\partial^2 S_B}{\partial U^2} < 0 \tag{12}$$

which says

$$1 - \beta U^{-1} \lesssim 0$$

or

$$M \gtrsim U \tag{13}$$

In other words (and seemingly by coincidence!) the known condition (13) for thermodynamic stability turns out to be equivalent (11) to the condition (10) that the radiation field act as a heat bath for the hole.

There are thus two limit cases to which Theorem 4 applies. In the first case, $M \gg U$, it merely requires the radiation field to be in a canonical state. But in the opposite limit, $U \gg M$, Theorem 4 would require the *hole* to be in a canonical state, which is impossible as already remarked. (To see this impossibility directly, notice that equation (1) implies

$$Z = \int_0^\infty e^{-\beta M} \frac{\partial N}{\partial M} dM \sim \int e^{M^2 - \beta M} dM$$

which diverges for all nonzero T.) The analysis of stability from the point of view of Theorem 4 has therefore (because of the aforementioned coincidence) merely confirmed the usual considerations in ruling out $U \gg M$.

On the other hand, the first case ($U \ll M$), where the black hole serves as a bath for the radiation field, would also seem to be impossible. For although Theorem 4 would not now require ρ_{hole} to be strictly canonical, the weaker requirement embodied in equation (7) (in which, recall, g_{jk} is periodic) is still enough to rule out a level density increasing as M^2. (The integral for Z would still diverge.) In view of the above discussion, this seems to mean that if in thermal equilibrium the state of the hole is not to depend on such features as the size and shape of the box enclosing it, then (insofar as one can even treat a black hole as a quantum system evolving in time according to Schrödinger's equation) there must be fundamental restrictions on the couplings possible between a black hole and its surroundings. If so, then to search for a restriction of the right sort might help one to understand the dynamics of black holes in general.

REFERENCES

Bratelli, O. (1978a) "Dynamics Stability and the KMS Condition in Quantum Statistical Mechanics," contributed to the Proceedings of the Conference on "Problemi Matematici nella Teoria dei Processi Quantistici Irreversibili," Laboratoria di Cibernetica del CNR, Arco Felice, Napoli, March 13–17.

Bratelli, O., Kishimoto, A., and Robinson, D. W. (1978b). "Stability and the KMS Condition," *Communications in Mathematical Physics*, **61**, 209–238.

Born, Max. (1949, 1964). *Natural Philosophy of Cause and Chance*. Oxford University Press, New York; Dover, New York.

Dixmier, J. (1969). *Les Algèbres d'opérateurs dans l'espace Hilbertian*. Gauthier-Villars, Paris.

Kato, T. (1966). *Perturbation Theory for Linear Operators*. Springer, Berlin.

Reed, M. and Simon, B. (1972). *Functional Analysis*, Vol. I. Academic Press, New York.

Riesz, F. and Sz.-Nagy, B. (1955). *Functional Analysis*. Frederick Ungar, New York.