Contents lists available at ScienceDirect

Studies in History and Philosophy of Modern Physics

journal homepage: www.elsevier.com/locate/shpsb

The case for black hole thermodynamics part I: Phenomenological thermodynamics

David Wallace

Dornsife College of Letters, Arts and Sciences, University of Southern California, United States

ARTICLE INFO

Article history:
Received 14 November 2017
Received in revised form
19 May 2018
Accepted 29 May 2018
Available online 18 June 2018

ABSTRACT

I give a fairly systematic and thorough presentation of the case for regarding black holes as thermodynamic systems in the fullest sense, aimed at readers with some familiarity with thermodynamics, quantum mechanics and general relativity but not presuming advanced knowledge of quantum gravity. I pay particular attention to (i) the availability in classical black hole thermodynamics of a well-defined notion of adiabatic intervention; (ii) the power of the membrane paradigm to make black hole thermodynamics precise and to extend it to local-equilibrium contexts; (iii) the central role of Hawking radiation in permitting black holes to be in thermal contact with one another; (iv) the wide range of routes by which Hawking radiation can be derived and its back-reaction on the black hole calculated; (v) the interpretation of Hawking radiation close to the black hole as a gravitationally bound thermal atmosphere. In an appendix I discuss recent criticisms of black hole thermodynamics by Dougherty and Callender. This paper confines its attention to the thermodynamics of black holes; a sequel will consider their statistical mechanics.

1. Introduction

Black hole thermodynamics (BHT) is perhaps the most striking and unexpected development in the theoretical physics of the last forty years. It combines the three main areas of 'fundamental' theoretical physics — quantum theory, general relativity, and thermal physics — and it offers a conceptual testing ground for quantum gravity that might be the nearest that field has to experimental evidence. Yet BHT itself relies almost entirely on theoretical arguments, and its most celebrated result — Hawking's argument that black holes emit radiation — has no direct empirical support and little prospect of getting any. So to outsiders — to physicists in other disciplines, or to philosophers of science — the community's confidence in BHT can seem surprising, or even suspicious. Can we really be so confident of anything without any grounding in observation?

In this article, and its sequel, I want to lay out as carefully and thoroughly as I can the theoretical evidence for BHT. It is written with the zeal of the convert: I began this project sharing at least some of the outsiders' skepticism, and became persuaded that the evidence is enormously strong both that black holes are thermodynamical systems in the fullest sense of the word, and that their thermodynamic behaviour has a statistical-mechanical underpinning in quantum gravity (and, as a consequence, that black hole evaporation is a unitary process not different in kind from the cooling of other hot systems, and that it involves no fundamental loss of information).

There are of course many reviews of this material. But those I know either (i) take for granted the main results of BHT, moving quickly over established material to get students up to speed with the research frontier; (ii) are explicitly historical, which illuminates how the community *in fact* came to accept BHT but can obscure the logic of whether and why they *should have* accepted it, or (iii) are written at a very high level of mathematical rigor, so high that a large fraction of the literature has to be omitted. I hope this paper will be complementary to extant material. With few exceptions, I present and describe results without going into the details of their derivation, and the student who wishes to properly understand the topic will need to read this paper in parallel with some of the extant review literature. My starting points (for this part of the paper) were Harlow (2016), Jacobson (1996, 2005), Thorne, Price, and Macdonald (1986), and Wald (1994, 2001).

A note on mathematical rigor: the tendency in foundational work on this subject (see, e. g., Belot, Earman, and Ruetsche (1999) and Earman (2011)) has been to work at the level of rigor typical in

E-mail address: dmwallac@usc.edu.

https://doi.org/10.1016/j.shpsb.2018.05.002
1355-2198/© 2018 Published by Elsevier Ltd.

mathematical physics, where all results are stated exactly and proved rigorously. This is much higher than the standard in theoretical physics more generally; it has the advantage of reliability, but the disadvantage that a very large fraction of the literature must be elided — especially in a frontier area like this, where the underlying physical principles are unclear and the mathematical framework partial and under active development. And the case for BHT — as will become apparent throughout this paper and, even more so, its sequel — rests not so much on individual results that have been established with full precision and rigor, but on the many independent calculations with different premises and approximation schemes that all lead to the same result. So this paper is written at the theoretical-physics level; I hope that readers who prefer their mathematics more precise will at least get a sense as to why *the community* takes BHT so seriously, even if they are not persuaded themselves.

This is a large topic, too large for any one paper. In this paper I confine my attention to phenomenological thermodynamics, setting aside any considerations of statistical-mechanical underpinnings for that thermodynamics. In Wallace (2017a) I consider the progress made in calculating the thermodynamical properties of black holes via statistical mechanics (in effective-field theory quantum gravity, in string theory, and via the AdS/CFT correspondence). And in Wallace (2017c) I use these two papers as a starting point to review and assess the notorious *information-loss paradox* which has motivated a large part of the critical attention paid to BHT.

The structure of the paper is as follows. I begin in section 2 by briefly reviewing classical thermodynamics, and discussing how it is modified for self-gravitating systems: to see whether black holes are thermodynamical, we need to be clear what thermodynamics is in the first place. In section 3 I consider classical black hole thermodynamics, arguing that while black holes offer a strikingly good realisation of the principles of thermodynamics when regarded as isolated systems, they completely fail to do so when considered as components of a larger system. In section 4 I show how including the implications of quantum field theory, in particular (though not exclusively) the Hawking effect, entirely remove this limitation; I also review the strength of the evidence for the Hawking effect itself, and the related but logically stronger claim that Hawking radiation leads to black hole evaporation. In an appendix, I address the arguments of a recent paper by Dougherty and Callender (2016) which criticises BHT (that paper was one trigger for my writing this paper, but engaging with its arguments in the main text would complicate my structure unhelpfully).

Readers familiar with extant debates on black hole thermodynamics may be surprised that in this paper I make virtually no mention of Bekenstein's classic argument (Bekenstein, 1973) for black hole entropy on the grounds of information. Partly this is because this paper is confined to phenomenological thermodynamics, and the relation of information to thermodynamics is normally made at the statistical-mechanical level. But mainly it is just because the link between information and thermodynamics is *controversial*, and so any argument for black hole thermodynamics from considerations of information is apt to inherit that controversy (recent critical takes on black hole thermodynamics by Wuthrich (2017) and Dougherty and Callender (*ibid*) both rely in one way or another on skepticism about the entropy-information link). Since (I will argue) we can make a compelling case for BHT and (in the sequel) black hole statistical mechanics without ever considering information, it seems simpler to sidestep the controversy. I discuss this in slightly more detail in the appendix.

I assume some familiarity with classical general relativity (in particular the Schwarzschild solution) and classical thermodynamics (and I quote standard results from both fields without explicit references); a little prior exposure to quantum field theory would also be helpful in section 4. Except where explicitly noted, I adopt units where $G = \hbar = c = k_B = 1$.

2. Thermodynamics and statistical mechanics: a brief review

Without any pretension to historical accuracy, complete precision or logical independence, we can break the salient parts of equilibrium thermodynamics into three: equilibrium and equilibration; the First and Second Laws for individual systems; interactions between multiple systems. (I believe the account I give basically tracks the consensus in mainstream physics; it broadly follows Wallace (2014, 2015).) I discuss each in turn; I then briefly consider the generalisation of equilibrium thermodynamics to *local* thermal equilibrium, and the subtleties introduced by gravitation. For this paper I do not need, and do not discuss, the statistical-mechanical underpinnings of thermodynamics.

2.1. Equilibrium and equilibration

A thermodynamic system has a family of *equilibrium states* parametrised by the energy and by a (usually small) number of additional conserved quantities and/or external constraints. In the absence of external interventions, if the system is in the equilibrium state corresponding to its constraints and conserved quantities, it remains in that state; if it is not, it *equilibrates*, evolving towards that state and reaching it, to any given degree of accuracy, after a finite time (Brown and Uffink (2001) refer to this equilibration principle as the *Minus First Law of Thermodynamics*).

For instance, for a box of gas (of some fixed kind of particle) the external constraint is the volume of the box, and the conserved quantities are the energy, the number of particles, and in principle the momentum and angular momentum. In general we assume a nonrotating box and study it in its rest frame, and/or assume that the box is so massive not to be affected by particle collisions, so that momentum and angular momentum may be neglected and 'energy' and 'internal energy' can be identified; often we also take the particle number as fixed and do not include it explicitly as a variable.

2.2. The First and Second Laws for individual systems

Given an isolated thermodynamic system, an *adiabatic* transformation of that system is some operation performed on the system, starting at equilibrium, that transforms its state to another equilibrium state without coupling it nontrivially to other thermodynamic systems.[1] Any such transformation can be thought of as a change to the external constraints and conserved quantities of the system via some external force; paradigm examples include expanding or compressing a gas, or putting a non-rotating system into rotation. The *work done* by such a process is defined as the change in the system's energy, and (by conservation of energy) is then equal to the energy cost to the external agent.

Only some such changes are physically possible by means of adiabatic transformations. Specifically, if the system's equilibrium states are parameterised by energy U and conserved quantities/external constraints X_i, there exists a function $S(U, X_1, \ldots X_N)$, called the *entropy* of the system (and hence defined, as far as thermodynamics is concerned, *only* at equilibrium), such that S is non-

[1] Many presentations introduce the notion of "heat" as a primitive, and define adiabatic processes as those that do not involve heat transfer; I do not do so here because it is convenient for BHT to treat heat as a derived quantity.

decreasing under any adiabatic transformation. This entropy non-decrease law is one form of the *Second Law of Thermodynamics*.

Adiabatic transformations can then be broken into three categories: *reversible* transformations, which leave S unchanged; *irreversible* transformations, which increase S, and *thermodynamically forbidden transformations*, which decrease S. It is generally the case that all reversible and irreversible transformations are physically performable (at least in principle, and perhaps in an idealised limiting case) so that the Second Law imposes a necessary and sufficient condition for a transformation to be possible. In particular, if we make a very small adiabatic change to the X_i and then wait for the system to re-equilibrate, that change will leave S unchanged to a very high degree of accuracy. So sufficiently slow adiabatic changes to the X_i will define processes which are very close to being reversible, becoming exactly reversible in the infinite-time limit. It is generally the case that such *quasi-static* transformations are always available.

We can express the entropy in differential form as

$$\mathrm{d}S = \beta\left(\mathrm{d}U + \sum_i \lambda_i X_i\right) \tag{1}$$

or, rearranging so that U is a function of S and the X_i,

$$\mathrm{d}U = T\mathrm{d}S - \sum_i \lambda_i X_i \tag{2}$$

where $T = 1/\beta$. T is called the *thermodynamic temperature* and the λ_i are the thermodynamic variables *conjugate* to the X_i; they can be given explicitly by

$$\frac{1}{T} = \left(\frac{\partial S}{\partial U}\right)_{X_i} \quad \lambda_i = T\left(\frac{\partial S}{\partial X_i}\right)_{U,X_j} \tag{3}$$

or by

$$T = \left(\frac{\partial U}{\partial S}\right)_{X_i} \quad \lambda_i = \left(\frac{\partial U}{\partial X_i}\right)_{S,X_j}. \tag{4}$$

The λ_i usually have a physical meaning: in particular, the variables conjugate to volume, momentum, angular momentum, particle number, and charge are, respectively, pressure, centre-of-mass velocity, angular velocity, chemical potential, and electric potential.

Equation (2) is one form of the *First Law of Thermodynamics*. It can be understood entirely statically, as a statement of the relations between different equilibrium states. But given the existence of quasi-static processes, we can also interpret it as describing the actual change in U induced by small adiabatic changes $X_i \rightarrow X_i + \delta X_i$ to the parameters, together with a flow of energy $Q = T\delta S$ into the system from some external reservoir. Following Wald (1994, p.141)) we can call these the *equilibrium-state* and *physical-process* interpretations, respectively. Flow of energy of this kind is called *heat flow* and makes sense even if the flow is not infinitesimal; conservation of energy entails that the change in a system's energy equals the heat flow into it plus the work done on it, which is another form of the First Law.

Finally, note that at this stage of our analysis S (and, hence, T) is fixed only up to an arbitrary rescaling: we can replace S with $f(S)$, for any smoothly increasing function f, and $1/T$ with $f'(1/T)$, without affecting anything said so far.

2.3. Multiple thermodynamic systems

Much of the content of thermodynamics is only available once we allow dynamical interactions between multiple systems. The rules for doing so are:

1. Any two systems may be placed in *thermal contact*, so that heat may flow between them while their other conserved quantities and external parameters remain *separately fixed.* This can be generalised to allow for other kinds of contact in which the two systems can exchange other conserved quantities.
2. Multiple systems in (perhaps-generalised) thermal contact may be treated as a single system; in particular, any such combined system will have a unique equilibrium state.
3. The Second Law of Thermodynamics generalises to require that the total entropy of two systems in (perhaps-generalised) thermal contact does not decrease when those systems exchange energy and other conserved quantities. For this to be well-defined, the possibility for rescaling of entropy decreases sharply: in multiple-system contexts, entropy must be taken as fixed up to a system-independent scale and a system-dependent additive constant.

From (2) and (3) together, it follows that:

4. If two systems are in thermal contact, and heat δQ flows from system 1 to system 2, the total change in entropy is $\delta S = \delta Q(1/T_2 - 1/T_1)$. So heat will flow only if $T_1 > T_2$, and indeed, no process can as its sole effect induce heat flow unless this condition holds (the *Clausius statement* of the Second Law). It follows that a necessary and sufficient condition for two systems in thermal contact to be jointly at equilibrium is that they are separately at equilibrium with equal temperatures. (This generalises to other forms of contact.) As a consequence, the relation 'at equilibrium with' is an equivalence relation: this is the *Zeroth Law of thermodynamics*, and in textbook presentations is often taken as a starting point; in my presentation, it is a consequence of other assumptions.
5. Given a process involving an infinitesimal heat flow between two equilibrium systems at thermodynamic temperatures T_1, T_2 together with work W done on the combined system, and such that the conserved quantities and external constraints of the two systems (other than energy) are unchanged at the end of the process, the First Law entails that

$$W = T_1\Delta S_1 + T_2\Delta S_2 = T_1\left(\Delta S_1 + \frac{T_2}{T_1}\Delta S_2\right). \tag{5}$$

Since the Second Law entails that $\Delta S_2 \geq -\Delta S_1$, we have

$$W \geq T_1\Delta S_1(1 - (T_2/T_1)). \tag{6}$$

From this, we can read off that the maximum efficiency of any cyclical process which generates work from heat flow between the two systems is $(1 - T_2/T_1)$ and, *a fortiori*, that no cyclical process can as its sole effect convert heat flow from an equilibrium system into work done, which is the *Kelvin statement* of the Second Law. (Other processes can do better, but they do not leave the other conserved quantities and constraints unchanged and so cannot be performed in a cycle.)

2.4. Local thermodynamic equilibrium

In an extended body (such as a solid, a fluid, or a field), if the rate at which a small region of the fluid equilibrates is fast compared to the rate at which it exchanges energy and other conserved quantities with neighboring regions, the body will approach *local thermal equilibrium*, at which we may express thermodynamic quantities like charge, energy, entropy, temperature and pressure as functions of position in the body. (For instance the sun, while not at equilibrium, is at local equilibrium, so that we can describe how temperature, pressure, entropy density and energy density vary from the core to the atmosphere.) Various phenomenological equations can be derived or postulated to describe the flow of thermodynamic quantities through the system. For instance, *Ohm's Law* describes how current flow in a conductor is dissipated as heat, and the *Navier-Stokes equations* describe the flow of a viscous fluid and the dissipation of organised energy as heat in that fluid. Various *transport coefficients*, like electrical resistivity and viscosity, appear in those equations, so that they cannot simply be derived from the equation of state but require additional empirical input.

2.5. Complications of gravity

Insofar as thermodynamics is the study of systems *at equilibrium*, it has fairly few real-world applications (except black holes themselves) to systems in which gravity is the dominant interaction. Indeed, a well-known result in celestial mechanics (the *gravothermal catastrophe* (Lynden-Bell and Wood (1968); Binney and Merrifield (1998), pp. 500–5) is a good introduction) demonstrates that classical Newtonian systems in which gravity is the only relevant form cannot reach stable equilibrium unless confined to a sufficiently small box. The only gravity-dominated astrophysical systems at thermal equilibrium (other than black holes themselves!) are degenerate-matter objects like white dwarfs and neutron stars, where quantum effects permit stability. Ordinary stars, for instance, are not at thermal equilibrium: there is a constant flow of energy from the core to the surface, and from the surface to interstellar space through emitted radiation; only the presence of fusion reactions in the core to replenish that lost heat allows stars to remain stable, until their fusion fuel is exhausted.

Nonetheless, thermodynamics can be coherently formulated for the artificial, but well-defined, example of (relativistic or Newtonian) self-gravitating systems confined to a box, where the existence of long-range forces in these systems leads to important subtleties, even before we consider black holes. Rather than discuss the (somewhat controversial) general structure of these subtleties (for that discussion, see Wallace (2010), Callender (2011), and references therein), I will illustrate them with a concrete example due to Sorkin, Wald, and Jiu (1981): a spherical box of radiation at thermal equilibrium, potentially large enough that self-gravitation has a discernible effect.

The sphere is assumed to be nonrotating and at rest. Its equation of state depends on two parameters: its radius R and its mass M (for a relativistic system, and in units where $c=1$, its mass and its energy can be identified). A crucial parameter is the *Schwarzschild radius* $R_S(M)=2GM/R$: if $R<R_S(M)$ then an event horizon forms around the sphere and it must be treated as a black hole.

As long as $R\gg R_S(M)$, gravitating effects are fairly insignificant and the sphere may be treated as if it were non-self-gravitating. It then behaves as a pretty conventional thermodynamic system, with an extensive equation of state determined by the intensive formulae

$$\rho=bT^4;\;\; s=\frac{4}{3}bT^3 \tag{7}$$

that determine the energy density ρ and entropy density s as functions of the temperature. In particular, if we consider a sequence of successively larger spheres with M/R^3 held constant, the temperature and density of each sphere likewise remain constant. But for denser spheres (the transition occurs roughly around $R\simeq 5R_S$) gravitational effects become highly important and the system displays several distinctive features characteristic of strongly self-gravitating systems (all discussed, or readily derived, in Sorkin et al.'s paper):

1. Because spacetime is nontrivially curved within the sphere, we cannot define the mass of the sphere simply as the integral of the local mass-density: indeed, that integral is not even well-defined in a coordinate-free way. Instead, the mass can defined by using Noether's theorem (according to which energy is the conserved quantity associated with time translation symmetry), calculated at a distance much larger than the shell radius at which the spacetime is approximately flat. The precise version of this concept of mass is called the *ADM mass*, after Arnowitt, Deser, and Misner (1962) (a related version, the *Bondi-Sachs mass* (Bondi, 1960, Sachs, 1961, 1962), is better suited to handle situations involving radiation but rests on the same basic idea). If the sphere had non-trivial spatial momentum and/or angular momentum, analogous ADM momenta and angular momenta can also be defined, using the appropriate asymptotic Noether symmetries.
2. The sphere becomes increasingly non-homogeneous, with the density being much higher towards the centre of the sphere. From this and the local equation of state (7), we can deduce that the locally-measured temperature also increases closer to the centre. The locally measured temperature $t(r)$ at a radius r from the centre is related to the thermodynamic temperature (given by $1/T=\partial S/\partial U$) by

$$t(r)=\alpha(m(r),r)^{-1}T \tag{8}$$

   where $m(r)$ is the mass of the sphere internal to r (more precisely: the ADM mass that the region of the sphere interior to r would have if it were confined to that region and the rest of the sphere removed) and $\alpha(m,r)=(1-2Gm/r)^{1/2}$ is the gravitational redshift induced by a spherically symmetric mass m.
3. The sphere is no longer extensive in any meaningful sense: increasing R to KR and M to K^3M will not produce a qualitatively similar sphere. Indeed, if $R<\sim 0.254R_S$, the sphere becomes unstable and undergoes gravitational collapse into a black hole.
4. The heat capacity of the sphere (i. e., the rate of change of mass with temperature at constant radius) decreases to zero and becomes negative, so that decreasing the energy of the sphere actually causes it to become hotter.

Though Sorkin et al. do not discuss it, the notion of “thermal contact” also has to be analysed with some care for these systems. For a start, we cannot put two such spheres in thermal contact simply by placing them adjacent to one another: their mutual gravitation would radically alter each other's states, probably producing gravitational collapse unless handled carefully. An intermediate system is required.

As a concrete example, consider the following process for transferring heat between two spheres with thermodynamic temperatures T_1,T_2, masses M_1, M_2 and surface redshifts α_1,α_2:

1. A box is slowly lowered to the surface of Sphere 1 from ‘infinity’ (i. e., from very far above the sphere), allowed to fill with a small amount of radiation of local mass m and temperature T_1/α_1, and then slowly lifted back to infinity, requiring (Unruh & Wald, 1982) work

$$W_1 = (1-\alpha_1)m. \tag{9}$$

2. The box is adiabatically compressed or expanded (as appropriate) to a temperature T_2/α_2, requiring additional (possibly negative) work

$$W_2 = \left(\frac{(T_2/\alpha_2)}{(T_1/\alpha_1)} - 1\right)m \tag{10}$$

   (as can be deduced from the equation of state (7)) and changing its mass to $m(T_2/\alpha_2)(T_1/\alpha_1)$.
3. The box is slowly lowered to the surface of Sphere 2, requiring negative work

$$W_3 = -(1-\alpha_2)\frac{(T_2/\alpha_2)}{(T_1/\alpha_1)}m. \tag{11}$$

4. The box is then opened and the radiation released into Sphere 2; this is adiabatic, since it has the same local temperature as Sphere 2's surface.

The entire process is adiabatic and has the following energy implications:

$$\Delta M_1 = -\alpha_1 m; \quad \Delta M_2 = \alpha_1 m(1+(T_2/T_1)); \quad W = W_1 + W_2 + W_3 \\ = \alpha_1 m(T_2/T_1). \tag{12}$$

This has the characteristic form of a Carnot cycle. As a corollary, if $T_1 > T_2$, net work is extracted by the process, and we can replace (3) by

3’. The box is slowly lowered towards the surface of Sphere 2 until the work extracted by doing so makes the whole process work-neutral, and then released to free-fall the rest of the way.

   The new process permits heat transfer, without work expenditure, from Sphere 1 to Sphere 2 provided $T_1 > T_2$, and so provides a means to put the two spheres in (somewhat indirect) thermal contact.

In many examples of self-gravitating bodies, there is another way to put two bodies into thermal contact: seal them both into a very large box with reflecting walls, and wait. If one or other body is above absolute zero, it will emit electromagnetic radiation; in due course, the box will fill with radiation in local thermal equilibrium. Each body is in thermal contact with the radiation and so, indirectly, with the other body. This is an effective way (in principle and in thought, not in engineering practice!) to, for instance, place two neutron stars or white dwarfs into thermal contact. It is not really an option for our radiation spheres, because they are themselves comprised of thermal radiation so the breakdown into subsystems would not be well-defined.

3. Classical black hole thermodynamics

We can now consider whether, and to what extent, these thermodynamic notions apply to black holes and systems of black holes. In this section I consider only ‘classical’ black holes, by which I mean: black holes, if we neglect or imagine away any quantum-field-theoretic effects: in particular, any matter fields present will be treated phenomenologically and classically. For clarity, I do not mean “black holes, under the fiction that the world is exactly classical”: I'm not sure that is even well-defined (though see Curiel (2014)) but in any case it presumably would not include thermal radiation, which can be treated phenomenologically as a classical fluid but whose derivation via statistical mechanics requires quantum theory.

3.1. *Black holes as objects*

The basic idea of BHT is that black holes are thermodynamic systems, and that a particular subclass of black holes (the stationary black holes) are the equilibrium states of those systems. But from the starting point of general relativity, it is hard to see how this is even coherent: in that context, a “black hole” is identified globally as a region of spacetime from which null geodesics cannot reach future infinity (see, e. g., Hawking and Ellis (1973)). A spacetime region cannot itself change in time, so the notions of ‘equilibrium’ or ‘equilibration’ don't obviously make sense under this definition.

But the relativist's concept of a black hole is not the only one extant in physics. *Astrophysicists* have long spoken of black holes as objects which persist through time and whose properties change in time: any talk of black holes orbiting one another, or of two black holes merging to form a larger hole, or of the velocity of a black hole relative to another astrophysical object, seems to require a three-dimensional view of black holes as objects, in tension with the spacetime-region view natural in theoretical relativity.

The *membrane paradigm* of Macdonald, Price and Thorne, developed in detail in the astrophysical context in Thorne et al. (1986) and adapted for the quantum theory of black holes by Susskind, Thorlacius, and Uglum (1993), addresses just this problem. Thorne et al. consider a timelike surface — the ‘membrane’, or ‘stretched horizon’ — that is placed around the true event horizon, at a very small proper distance from the true horizon. Thorne et al. give the stretched horizon an area $(1+\alpha)^2$ times that of the true horizon, where α is some positive real number $\ll 1$; more useful for foundational purposes is Susskind et al.'s convention (which I adopt henceforth), giving the horizon an area one Planck area larger than that of the true horizon.

The defining property of the event horizon, physically, is that nothing can emerge from it, and so in particular nothing can enter it and later return. But (in an admittedly somewhat heuristic sense), the stretched horizon is so close to the true horizon that *virtually* nothing can cross the stretched horizon and return, because doing so would require extremely high accelerations (for timelike bodies dropped into the hole) or extremely short wavelengths (for photons, or extremely relativistic massive particles, on trajectories that pass between the stretched and true horizon) — indeed, under Susskind et al.'s convention, it would require accelerations so high, and/or wavelengths so short, as to require Planck-scale physics to describe.[2] So as long as we are dealing with energy levels well below the Planck scale, the stretched horizon may be treated as a one-way barrier just as can the true horizon.

On the other hand, the stretched horizon is an ordinary timelike surface; it can be treated as a two-dimensional closed surface in

[2] Thanks to Erik Curiel for pointing out the short-wavelength photon case.

space that evolves through time, and so can be attributed potentially-time-dependent physical properties. And with its aid, we can then restate the goal of black hole thermodynamics as follows: to investigate the extent to which the stretched horizons of black holes can be treated as ordinary physical systems, and assigned mechanical, electromagnetic, and thermodynamic properties, from the point of view of any observers who remain outside the black hole — or, to put it in less operational terms, the extent to which we can give a self-contained account of physics in the region of spacetime exterior to any black holes in terms of stretched horizons to which such properties are assigned.

3.2. *Equilibrium and equilibration for black holes*

Thermodynamics describes equilibrium systems in terms of their conserved quantities and external constraints. There are no real external constraints applicable to a black hole, but there are quantities which we would expect to be conserved: the energy, momentum and angular momentum of the hole (defined asymptotically by the ADM method) and its electrical charge. In each case these quantities are associated to long-range forces (gravity for the quantities associated to spacetime symmetries; electromagnetism for charge), as these forces ensure that matter bearing the conserved quantity will leave an asymptotic trace on the spacetime even once it crosses the stretched horizon. (Conserved quantities like baryon number, by contrast, cannot be expected to show up in the physics of the black hole exterior, since the long-range physics will be indifferent as to whether a particle that crosses the horizon is, say, a neutron rather than an anti-neutron.) By working in a reference frame at which the black hole is at rest and its angular momentum is aligned along the z axis (again, using the ADM charges to define this rigorously) we reduce the conserved quantities to three: the black hole's mass M, the magnitude J of its angular momentum, and its charge Q. So if black holes have equilibrium states, we would expect the space of such states to be parametrised by these three quantities.

The definition of an 'equilibrium' state is that it is unchanging in time, and general relativity offers a clear way to represent this: we look for *stationary* solutions of the Einstein field equations, that is: solutions with a timelike Killing vector. Such solutions certainly exist for general M, J, Q: the *Kerr-Newman* solutions to the coupled equations of general relativity and vacuum electromagnetism (aka Einstein-Maxwell theory) are stationary and parametrised precisely by mass, angular momentum and charge. When $Q = 0$, these solutions reduce to the Kerr solutions of vacuum general relativity; when $J = 0$, to the spherically-symmetric Reissner-Nordstrom solutions of the Einstein-Maxwell theory; when both are zero, to the well-known Schwarzschild solution. The Kerr-Newman solution only describes a black hole when $Q^2 + J^2/M^2 \leq M^2$, with solutions violating this inequality describing naked singularities; black holes that saturate the inequality are called *extremal*, and are a somewhat puzzling special case (one that has been of considerable importance in quantum gravity, as I discuss in the sequel to this paper).

The 1970s saw extensive work by Bardeen, Carter, Hawking, Israel and many others to prove the "No-Hair Conjecture": that the Kerr-Newman black holes are the unique stationary solutions to the Einstein-Maxwell theory, and so provide unique equilibria. To this day there remain loose ends in the conjecture and in its extension to more general situations in higher spacetime dimensions and with other long-range forces present, but in his review article in the Einstein Centenary Survey Carter (1979) felt able to say that

> the no-hair theorems available … are quite sufficient to justify — with at least the degree of rigour usually considered acceptable in physics — the assumption by any practically minded astrophysical theorist that any (external source free) black hole equilibrium-state solution … belongs to the Kerr or Kerr-Newman families".

(See Carter's review article for detailed references and for a summary of the main results; see also Carter (1997) for some historical remarks and Chrusciel and Costa (2008) for a fairly up-to-date survey.)

Of course, *thermodynamic* equilibrium requires more than mere stationarity: it requires *non*-equilibrium systems to converge to equilibrium, and in particular, perturbations of equilibrium states to be damped back down to equilibrium. The stability of black holes, and the convergence to equilibrium of non-stationary black holes, has been extensively studied both analytically and numerically. By the mid-1980s (see chapters VI-VII of Thorne et al. (1986), and references therein) it was established that perturbations of the stretched horizon by external gravitating bodies are damped away (for instance, the stretched horizon can oscillate, but these oscillations are damped, dying away back to equilibrium via the emission of gravity waves). Computer simulations of colliding black holes, and accretion of matter onto black holes, likewise demonstrate that the system evolves rapidly to the equilibrium-black-hole configuration, decaying by the emission of gravity waves ('ringdown'). And the historic observation of gravity waves in 2016 by the LIGO observatory (Abbott and Collaboration) 2016) provided a remarkably precise fit to the quantitative ringdown predictions, and so can reasonably be said to provide (ongoing) *observational* support for black hole equilibration.

In summary: we have both a clear understanding of *what* the black hole equilibria are, and a pretty good grasp on *why* they are indeed equilibria: at the least, I think it would be hard to argue that we have any better theoretical control of how paradigm 'normal' thermodynamical systems, like dilute gases, approach and remain at equilibrium. So far, black holes fully fit the requirements to count as thermodynamic systems.

3.3. *The laws of black hole thermodynamics*

To treat a black hole as a thermodynamic system requires us to identify external interventions, and to divide them into adiabatic changes and heat flows. The former is fairly straightforward: to move a black hole from one equilibrium state to another is going to require us to change its mass, angular momentum or charge, and the simplest way to do that is to drop matter into it. The latter is more delicate, since the division between 'heat' and 'work' is less obvious in an alien situation like this than for a box of gas. The simplest thing to do (in this case as in other less-familiar cases in 'regular' thermodynamics) is to identify which transformations are reversible and which irreversible, and then define the quasi-static adiabatic processes as the reversible ones.

Christodolou and Ruffini demonstrated (Christodolou and Ruffini (1971); see Misner, Thorne, and Wheeler (1973, pp. 907–913) for a discussion) that the quantity that plays the role of entropy for a black hole (at least for infinitesimal changes) is surface area (which, for an equilibrium black hole, is given by a known function of M, J and Q): any intervention on an equilibrium black hole must leave the surface area nondecreasing, so that the reversible processes are those that leave surface area invariant and the irreversible processes strictly increase area. Reversible transformations of J and Q can be brought about as follows:

- To reversibly change the charge of a charged black hole, lower some charged matter very slowly on a cord so that it is suspended, stationary, just above the event horizon; then let go.

- To reversibly increase the angular momentum of a rotating black hole, fire some mass at it on a trajectory which just brushes the event horizon.
- To reversibly decrease the angular momentum of a rotating black hole, use the *Penrose process* (Penrose (1969), Penrose and Floyd (1971); see Carroll (2003, pp. 267–271) for an introduction): fall freely towards the black hole on a trajectory that passes just above the event horizon, and at point of closest approach, eject some mass into the black hole on a trajectory opposite to the direction of rotation of the hole.

Dropping charge into a black hole from finite height, or injecting mass on a non-brushing trajectory, or using the Penrose process on a higher trajectory, will in each case be *irreversible*, bringing about an increase in surface area.

Hawking's area theorem (Hawking, 1972) generalises Cristodolou and Ruffini's result beyond infinitesimal changes: Hawking proved that the area of any black hole is nondecreasing. His derivation presumes

1. that physics in the exterior of the black hole remains predictable (that is, roughly: assuming that no naked singularities form; see Wald (1994, pp.138–9) for a more precise discussion);
2. the *null energy condition*: that the stress-energy tensor T satisfies $T(v,v) \geq 0$ for any null v. This is violated in some exotic quantum-field-theoretic situations (of which more later) but seems a safe assumption for bulk matter, such as electromagnetic radiation and astrophysical fluids.
3. the Einstein field equations, which translate the null energy condition into a condition on the Einstein tensor. (Once that translation is made, the area theorem is purely a result in differential geometry, with no additional dynamical input.)

Bardeen, Carter, and Hawking (1973) christened the Area Theorem the "Second Law of black hole thermodynamics"; in fact, it goes rather beyond the entropy-increase form of the standard Second Law, since black hole surface area remains well-defined even when a black hole is far from equilibrium, whereas thermodynamic entropy is defined only at equilibrium.

In the same paper, Bardeen et al. also established the "First Law of black hole thermodynamics" which states that

$$\mathrm{d}M = \frac{1}{8\pi}\kappa \mathrm{d}A - \Omega \mathrm{d}J - \Phi \mathrm{d}Q \tag{13}$$

where κ is the surface gravity of the black hole, A its surface area, Ω its angular velocity, and Φ the electric potential on its surface. This is *precisely* the form of the standard First Law for a thermodynamic system where angular momentum and charge are conserved quantities, including the identification of the conjugates to J and Q as, respectively, angular velocity and electric potential. It permits us to identify the thermodynamic temperature of the hole as proportional to the surface gravity — albeit, as long as we are considering a system in isolation, we have only identified entropy up to a monotonic function. Furthermore, we can independently prove the 'physical-process' and 'equilibrium-state' versions of the First Law distinguished by Wald (recall the discussion in section 2.2), demonstrating that the overall structure of interventions on the black hole is self-consistent and fits the model of equilibrium thermodynamics.

3.4. *Beyond Einstein's equation*

Bardeen et al.'s derivation of the laws of Black Hole thermodynamics presupposed the Einstein field equations; however, as Wald and collaborators have shown (Wald (1993); see Wald (1994, pp.143–147) for an introduction and further references, and Jacobson and Mohd (2015) for more recent developments), the First Law (in both physical-process and equilibrium-state form) can be derived from a general diffeomorphism-invariant Lagrangian theory of gravity by identifying the entropy as (a form of) the Noether charge associated with the diffeomorphism symmetry, evaluated with respect to a vector field that coincides on the horizon with the horizon Killing vector.

So far as I know there is no fully general non-decrease theorem for this generalised black hole entropy of the same scope of Hawking's area theorem, but Jacobson, Kang, and Myers (1995) have demonstrated that this generalised definition of entropy is nondecreasing under at least quasi-stationary processes, provided that the null energy condition is satisfied; they also prove the analog of Hawking's result for a large class of generalisations of the Einstein Lagrangian.

The physical reason for caring about this generalisation lies in the effective-field-theory program in contemporary particle physics. From that perspective, general relativity is thought of as a non-renormalisable effective field theory, regularised by a cutoff imposed by unknown Planck-level physics. In such a theory, all possible diffeomorphism-covariant action terms should be present; the Einstein-Hilbert action is just the leading-order term in an infinite expansion of the Lagrangian in these various terms. So the fact that black hole thermodynamics extends so naturally beyond the Einstein-Hilbert case is reassuring for the physical applicability of the theory.

3.5. *Local properties of the stretched horizon*

The stretched horizon of a black hole is, it seems, a purely *fictional* entity, invisible to anyone falling through it and corresponding to no locally-present distribution of charge or energy. It is therefore frankly startling that it can be treated not simply as a formal device to make sense of black hole thermodynamics (as I used it above) but as an actual extended physical system with local thermodynamic properties.

To expand: as discussed *in extenso* in Thorne et al. (1986) and references therein, we can treat the stretched horizon as a two-dimensional, electrically-conducting, viscous fluid, assigning to each infinitesimal part of its surface the exact charge, current, and stress-energy densities required to terminate the electromagnetic and gravitational field lines on its exterior. This assignment is arguably fictional since an observer freely falling through the horizon will not encounter these charges or energies, but from the point of view of physics outside the stretched horizon they are entirely real. To give some examples (many more can be found in Thorne et al.):

1. If a positively charged particle falls towards the North pole of an uncharged black hole, its field will induce a current flow of negative charge towards the north pole, which will become negatively charged; the South pole, opposite the direction of approach of the falling particle, will become positively charged. By applying the law of Ohmic dissipation to this current flow (the black hole's surface resistivity is $\sim 377\,\Omega$) we deduce that heat will be dissipated in this process so that the black hole area increases. When the charged particle reaches the surface, current will flow back until the charge density on the surface is constant, dissipating more heat. Any region of charge excess will spread out exponentially so that the time for an initially non-equilibrium charge distribution to equilibrate is $\tau_{eq} =\sim M \log M$ in Planck units, or in more astrophysically useful units

$$\tau_{eq} \sim 4.9 \times 10^{-6}\left(\frac{M}{M_\odot}\right)(\log(M/M_\odot)+87.4)\text{seconds} \qquad (14)$$

($M_\odot = 1.99 \times 10^{30}$kg is the mass of the Sun). Only in the limit where the charge is lowered infinitely slowly to the surface will the current flow be so slow, and the readjustment of charge across the surface so complete, that no heat is dissipated; this is the reversible process described previously. (Znajek, 1978; Damour, 1978; Macdonald & Suen, 1985; Thorne, Price, and Macdonald 1986, pp. 35–38,57–64.).

2. If an electrically neutral black hole rotates in an asymptotically constant magnetic field at right angles to its axis of rotation, eddy currents will be induced in the horizon. The magnetic field will exert a torque on the black hole via these currents, which will slow its rotation while also dissipating heat through electrical resistance. The result is that the rotational energy of the black hole will be dissipated as heat, slowing the black hole's rotation and increasing its area; the overall energy of the black hole remains conserved: that is, no energy is extracted from the static magnetic field in this process (Thorne & Macdonald, 1982; Thorne, Price, and Macdonald, 1986, pp. 102–106).
3. If a black hole rotates in the tidal field of a larger gravitating body, the surface of the hole will be perturbed; this in turn produces viscous dissipation and corresponding viscous torque on the black hole in accord with the Navier-Stokes equation, dissipating heat and slowing the rotation of the hole (Hawking & Hartle, 1972; Hartle, 1973, 1974; Thorne, Price, and Macdonald, 1986, pp. 252–255.).

   Also part of the local thermodynamics of black holes is the so-called *Zeroth law of black hole thermodynamics* (Bardeen et al., 1973), which states that the temperature of a black hole is constant everywhere on the horizon. In ordinary thermodynamics, the analogous result — that for a body at equilibrium, the local temperature is constant — is more naturally thought of as a corollary of the Zeroth Law applied to the local-thermal-equilibrium context.

3.6. *No thermal contact for classical black holes*

So far as we treat each black hole as an isolated system, the resemblance to a thermodynamic system seems pretty complete: black holes have notions of equilibrium and equilibration, reversibility and irreversibility, and local thermodynamic properties. But the resemblance terminates abruptly — at least as far as classical black holes are concerned — as soon as we try to consider them as thermodynamic systems interacting with other black holes, or with non-black-hole thermodynamic systems.

Specifically: there seems to be no available process that can reduce the entropy of one black hole and increase that of another (or of a non-black hole thermodynamic system), even if the total entropy is increasing. To the contrary, the analysis of reversible and irreversible processes above applied to each hole separately. Likewise, Hawking's area theorem applies separately to each connected component of a spacetime's event horizon, and so mandates not just that the total entropy of a system of black holes is nondecreasing but that the entropy of each black hole is separately nondecreasing. As a corollary, there seems no prospect of running a Carnot cycle between two black holes, and no prospect of allowing heat to flow from one hole to another. Likewise, there seems no way to make sense of heat flow from a black hole, to any other thermodynamic system. The nearest we can get is to allow two black holes to 'interact' by colliding, in which case the area theorem guarantees that the new black hole has a larger entropy than its constituents, but this is a pale shadow of genuine thermal contact.

In particular, classical black holes are completely black in the sense that they omit no thermal radiation. This means that a black hole placed in thermal contact with another body by the method of putting both in a box and letting it fill with radiation will simply eat all the radiation, however low its temperature. The only temperature that we seem consistently able to attribute to a classical black hole is then absolute zero.

These limitations are aggravated by Bekenstein's (1973) observation that identifying black hole area with entropy also provides opportunities to violate the Second Law of thermodynamics unless we place some constraints on the form of the energy-entropy relation for ordinary matter — constraints that do not seem well motivated within classical physics. Specifically:

- If some body of small mass m and entropy s is slowly lowered right to the event horizon and then released (the so-called 'Geroch process', proposed by Robert Geroch during a 1970 Princeton colloquium), it will do work on the mechanism that lowers it. Qualitatively this is no different from the way in which a weight slowly lowered from a pulley can do work at the top of the pulley, but the quantitative scale is much larger: if a point-like body of mass m is slowly lowered to a point above the event horizon with redshift α, then the work extracted is $W = m(1-\alpha)$ and so (by conservation of ADM mass) the mass increase of the black hole is $m\alpha$ (Unruh & Wald, 1982). As the mass is lowered arbitrarily close to the horizon, $\alpha \to 0$, and so the black hole's mass after the process is carried out, and hence its surface area, will be unchanged — but the entropy of the outside world will decrease by s. (This process can even be used to turn heat into work with perfect efficiency, thus violating at least the operational content of the Kelvin statement of the Second Law.)
- If some large body with mass M and entropy S undergoes gravitational collapse, it will form a black hole with area proportional to M^2, and decrease the entropy of the external world by S. If black hole area is identified with entropy (up to some scale factor K) then the total entropy change is $16\pi KM^2 - S$, which for appropriate choices of M and S could easily be negative (Susskind, 1995).

As Bekenstein pointed out, both of these arguments would fail if there is some fundamental bound on the minimum size of a body with given entropy and mass. To expand: for simplicity let us specialise to a Schwarzschild (i. e., nonrotating, uncharged) black hole, where the metric is

$$\mathrm{d}s^2 = -\alpha(r)^2\mathrm{d}t^2 + \alpha(r)^{-2}\mathrm{d}r^2 + r^2\left(\mathrm{d}\theta^2 + \sin^2(\theta)\mathrm{d}\phi^2\right) \qquad (15)$$

with $\alpha r = \sqrt{1-2GM/r}$. $\alpha(r)$ can be interpreted as the 'redshift' at radial coordinate r, i. e. the time dilation, relative to clocks at infinity, measured by an observer hovering above the black hole at constant radial coordinate r. The proper distance from the event horizon of an object at coordinate r is

$$d = \int_{2M}^{r} \frac{\mathrm{d}r}{\alpha(r)} \qquad (16)$$

Very close to the black hole ($(r-2M)/2M \ll 1$), we can approximately take

$$\alpha(r) \simeq \left(\frac{r-2M}{2M}\right)^{1/2}, \tag{17}$$

evaluate d, and solve to get

$$\alpha(d) = d/M. \tag{18}$$

So a spherical body of radius d, entropy s, and mass m, lowered slowly into the black hole, will increase the mass of the black hole by $\delta M = md/M$, and so the black hole entropy by $\delta S_h = 8\pi M \delta M = 8\pi md$. The total increase in (black hole entropy plus outside-matter entropy) is then

$$\Delta S = \delta S_h - s = 8\pi md - s. \tag{19}$$

If some new principle of nature means that any such body must satisfy $s/m \leq 8\pi d$, that would suffice to ensure $\Delta S \geq 0$ (changing the geometry of the body changes the numerical coefficients but not the overall argument). A qualitatively similar constraint, $s/m \leq 2\pi d$, also blocks Susskind's argument from gravitational collapse: the body, on forming a black hole, will have entropy $S_h = 4\pi m^2$, so the net increase in entropy is

$$\Delta S = 4\pi m^2 - s = 2\pi m(2m - s/2\pi m) \geq 2\pi m(2m - d). \tag{20}$$

But the body must initially lie outside its own Schwarzschild radius, $d > 2m$, to have avoided collapse already, so this must be positive.

However suggestive this *Bekenstein bound* might be, however, there is at least within classical physics no obvious reason why it must hold. And so to sum up: although classical black holes have some highly thermodynamic-*like* properties, core aspects of thermodynamics depend on interactions between thermodynamic systems; these interactions do not seem to function correctly for classical black holes, rendering the analogy with thermodynamics purely formal.[3]

4. Quantum field theory

Quantum mechanics — specifically, quantum field theory, formulated on a classical but curved spacetime — removes the blemishes in BHT and transforms it from a suggestive analogy to a full equivalence. The central result here is the *Hawking effect*: the discovery that black holes emit thermal radiation, at exactly the temperature that BHT would predict.

4.1. Hawking radiation

In this section I want to simply state what Hawking radiation is, and give some insight into its properties, leaving the question of whether we should believe it exists to the next section. As a starting point to understand Hawking radiation, let's consider for simplicity a free, massless, scalar quantum field theory defined on Schwarzschild spacetime, with metric (15). (Throughout this section, I assume a ’large’ black hole, where curvatures outside the black hole are small compared to the Planck scale and hence quantum-gravitational effects can be neglected; the description of black hole radiation from Planck-scale black holes lies beyond currently-understood physics.) The ‘external’ region of that spacetime — the region outside the event horizon, defined by $r > 2GM$ — is a globally hyperbolic spacetime suitable for describing the exterior of an uncharged non-rotating black hole. Since it has a timelike Killing vector — corresponding to translation in the t coordinate — we can coherently analyse the eigenstates of energy of the theory, and since the field is free, those eigenstates can be defined by the occupation number of the various independent modes of the field, which are the definite-frequency solutions of the Klein-Gordon equation on the Schwarzschild background.

Given the linearity of the Klein-Gordon equation, and given the time-translation and rotational symmetries of Schwarzschild spacetime, any solution of the Klein-Gordon equation can be written (here I follow Harlow (2016, pp.27–29)) in the form

$$\Psi(t, r, \theta, \phi) = \int \mathrm{d}\omega \sum_{l,m} \alpha_{l,m}(\omega) f_{\omega lm}(t, r, \theta, \phi) \tag{21}$$

where

$$f_{\omega lm}(t, r, \theta, \phi) = \frac{1}{r} Y_{lm}(\theta, \phi) \mathrm{e}^{-i\omega t} \psi_{\omega l}(r) \tag{22}$$

and Y_{lm} is a spherical harmonic.[4] All of the detailed physics of the wave equation is contained in the functions $\psi_{\omega l}(r)$, with the rest following purely from the symmetry structure of the theory (recall that solutions to the Schrödinger equation for a Coulomb potential, for instance, have the same form). So to understand the solutions, we need to understand the features of these functions.

To describe them further, it is helpful to introduce the *tortoise coordinate* r_*, defined by

$$r_* = r + \ln|r/2GM - 1|, \tag{23}$$

which approximates r for $r \gg 2GM$ but stretches the distance to the event horizon to cover the whole negative-x axis; it also simplifies matters to adopt, temporarily, units in which $2GM = 1$, i. e. to use the Schwarzschild radius as our unit of distance. The radial function $\psi_{\omega l}$ then satisfies

$$\left(-\frac{\mathrm{d}^2}{\mathrm{d}r_*^2} + V(r)\right)\psi_{\omega l} = \omega^2 \psi_{\omega l} \tag{24}$$

where

$$V(r) = \frac{r-1}{r^3}\left(l(l+1) + \frac{1}{r}\right) \tag{25}$$

and where r is given implicitly in terms of r_* by (23).

Formally (24) is just the nonrelativistic Schrödinger equation in one dimension, so that the problem of solving the Klein-Gordon equation has been reduced to a scattering problem in one dimension. Modes can be thought of as incoming either from infinity or from the event horizon, and they will scatter off, or tunnel through, a potential barrier whose form depends on the angular momentum l. For $l \gg 1$ the barrier has height $\sim l^2$ and is located at $r = 3/2$.

We can now distinguish (following (Thorne et al., 1986, ch. VIII)):

- IN modes, which come in from infinity and largely scatter off the angular-momentum barrier (for $l \gg 1$) with some small amplitude to penetrate the barrier and fall onto the event horizon;
- UP modes, which come up from the vicinity of the event horizon and are largely trapped close to the horizon by the angular-

[3] Curiel (2014) challenges this result and argues for a fully thermodynamic understanding of black holes even in the classical case; engagement with these arguments lies beyond the scope of this paper.

[4] For a reminder of the properties of spherical harmonics, see, e. g., Jackson (1999).

momentum barrier (for $l \gg 1$) with some small amplitude to escape to infinity.

Hawking's result is then the following: for a black hole formed by gravitational collapse, and with surface gravity κ, the quantum state of the exterior is a thermal state with respect to the UP modes, at a temperature $\kappa/2\pi$. (With respect to the IN modes, the quantum state is determined by boundary conditions; for an astrophysical black hole in the current epoch, for instance, we might take the IN modes to be in a thermal state at the temperature of the microwave background radiation.)

To understand this summary, it is helpful to describe the radiation as seen by a fictional observer hovering at a fixed distance above the black hole. Such observers move along a trajectory of constant r, θ, ϕ, and are often called *fiducial observers*, or FIDOs. A fiducial observer at a redshift of α follows an accelerated worldline with locally-measured acceleration $\alpha^{-1}\mathrm{d}\alpha/\mathrm{d}r$; we can imagine the observer being held in place by a rope supported at infinity. Fiducial observers observe very different effects depending on how close they are to the event horizon:

- A fiducial observer close to the event horizon (i. e., whose distance to the event horizon is small compared to the Schwarzschild radius, and in particular who is between the potential barrier described by equation (25) and the event horizon) observes a thermal bath of black-body radiation which might be thought of as the black hole's "atmosphere": this radiation remains largely trapped by the potential barrier and mostly falls back into the black hole rather than escaping to infinity. The apparent temperature of the radiation, as measured by the fiducial observer, will be $T/\alpha = \kappa/2\pi\alpha$, because that observer's clocks are redshifted by a factor α compared to coordinate time; we have already seen that this shifting of the temperature is a general feature of self-gravitating thermal systems.
- When the observer's redshift is large enough that the locally measured temperature approaches the Planck temperature, the field-theory model we have used becomes unreliable: put another way, at this redshift the locally-measured wavelength of the radiation approaches the Planck length and we expect quantum-gravitational effects to cut off the QFT description. This occurs (not by coincidence) when the fiducial observer has reached the *stretched* horizon, using Susskind et al.'s convention for its location.
- Conversely, an observer *far* from the event horizon sees a stream of outwardly flowing radiation appearing to emerge from the black hole. This radiation is *not* black-body radiation, because modes of different angular momentum escape the black hole atmosphere to differing degrees. The *grey-body factors* of a black hole describe how the black hole's emission spectrum, as a function of angular momentum and frequency, deviates from a perfect black body.

There is also a divergence between the observations of *fiducial* observers, and those of *inertial* observers falling into the black hole from far away, a divergence which increases as the event horizon is approached. In the outer (radiation) region of the spacetime, both groups of observers have similar experiences: they see an outward-going stream of radiation (although the increasing velocity of the infalling observer, and increasing acceleration of the fiducial observer, cause these experiences to diverge increasingly as they approach the black hole). Within the black hole atmosphere, and particularly as the observers approach the stretched horizon, the experiences become sharply different: while the fiducial observers experience ever-hotter thermal radiation, the infalling observer sees only slight deviations from empty spacetime.

4.2. Evidence for the Hawking effect

Before considering the thermodynamics of black holes in the light of Hawking radiation, we should pause briefly to ask how confident we should be in its existence. After all, while the classical theory of black holes lies within the range of astrophysical observation and so is supported by quite a lot of direct evidence, there is no realistic prospect of observing Hawking radiation from astrophysical black holes, and so far no proposal for observing it in non-astrophysical contexts (e. g. at the LHC, or through the decay of primordial black holes) has borne fruit. So the case is entirely theoretical; it is, nonetheless, very powerful.

To my knowledge there are at least five independent, conceptually distinct routes by which the Hawking effect can be derived:

1. Hawking's original method of matching outgoing modes with exterior modes via the technique of Bogoliubov transformations (Hawking (1975); see Wald (1994, ch.7) for a review);
2. Making precise the heuristic understanding of black hole evaporation by particles tunneling across the event horizon (Parikh & Wilczek, 2000);
3. Requiring the quantum state of the black hole exterior to solve or nearly solve the semiclassical Einstein field equations, which is possible only if the outgoing modes are in a thermal state at the correct temperature (Candelas (1980), Sciama, Candelas, and Deutsch (1981); see also section 4.3);
4. Path-integral methods on the analytic continuation of the black hole exterior spacetime, which demonstrate that the radiation-free vacuum — and, more generally, any thermal state at the wrong temperature — leads to singularities at the horizon (Hartle & Hawking, 1976; Israel, 1976);
5. Observing that radiation flow across the event horizon is necessary to prevent anomalous breaking of the diffeomorphism symmetry (Robinson & Wilczek, 2005).

Each has its strengths, weaknesses, and distinctive features. Hawking's original approach (1) is perhaps most directly tied to the physics of actual collapse-formed black holes, but is confined to free fields. At the other extreme, (4) is completely general but only applies to a black hole at thermal equilibrium with an external radiation bath, requiring additional physical justifications to be applied to collapse-formed black holes. (1) and (2) give concrete mechanisms for Hawking radiation, whereas (3)–(5) derive contradiction or unphysical paradox from its absence. But collectively, they strongly suggest that Hawking radiation really is a consequence of quantum field theory on curved spacetime, and not simply an artefact of a particular method of mathematical analysis. In turn, quantum field theory on curved spacetime is just an application of the general machinery of modern quantum field theory (in particular, the use of field theory to describe quantum fluctuations against a fixed classical background) and — while it is fair to note that it has not passed the sort of precision tests which underpin support for, say, flat-space quantum electrodynamics, and that the notorious cosmological-constant problem[5] gives some grounds for concern about its overall coherence — it is the theoretical underpinning for experimentally-tested results in astrophysics and cosmology, notably interferometry experiments involving photons that have passed through regions of curved spacetime.[6]

[5] See Weinberg (1989), Peebles and Ratra (2003), and references therein; thanks to Erik Curiel for pointing out its significance here.

[6] Thanks to an anonymous referee for pressing the question of what grounds we have to accept curved-spacetime QFT.

It is also possible to give a fairly direct *physical* argument for Hawking radiation. Consider a fiducial observer, very close to the event horizon (at some redshift $\alpha \ll 1$, say). The radius of curvature of the spacetime is much larger than the distance to the horizon, so locally it will appear to the observer as if they are accelerating in flat space at a constant locally-measured acceleration $\alpha^{-1}\mathrm{d}\alpha/\mathrm{d}r$. Sufficiently close to the event horizon, this tends to κ/α, where κ is the surface gravity. The Unruh effect (Unruh (1976); see Harlow (2016, pp.15–24) for a helpful discussion, and Crispino, Higuchi, and Matsas (2008) for an exhaustive review) tells us that an observer in flat spacetime with uniform acceleration a experiences a bath of black-body thermal radiation at a temperature of $a/2\pi$ (and the Unruh effect itself can also be derived in multiple ways: from Bogoliubov methods, via path integrals, and as a rigorous result in algebraic quantum field theory, to name three). So by the equivalence principle, we would expect our fiducial observer to see something very close to thermal radiation at this temperature: that is, at locally measured temperature $\kappa/2\pi\alpha$.

Now, very close to the black hole the event horizon fills almost the whole sky, so we would expect most of the radiation observed by the fiducial observer to fall back into the black hole. But it doesn't *quite* fill the whole sky, so any given radiation mode will have some amplitude to escape to infinity (with lower-angular-momentum modes having the highest amplitude). That radiation will be redshifted by a factor α and so will be seen at infinity to have a temperature $\kappa/2\pi$, in accordance with Hawking's prediction (and to be radially streaming from the black hole).

I pause to consider and rebut a well-known potential objection to the existence of Hawking radiation: the so-called *trans-Planckian problem*. In a nutshell, the problem is that radiation observed from a black hole sufficiently long after it forms is apparently redshifted down from radiation at a locally-measured wavelength shorter than the Planck length, i. e. a wavelength at which we should regard quantum field theory as unreliable in any case. (And "sufficiently long" is not at all long, in astrophysical terms: the timescale is $\sim M \log M$ in Planck units, or (from equation (14)) $\sim 10^{-3}$ seconds for solar-mass black holes.) At times much later than this, the original energy of the detected radiation gets *bigger* than Planckian, indeed ridiculously big.

If this argument were correct, it would demonstrate not simply that Hawking radiation is absent, but that there is some inherent inconsistency in defining quantum field theory on a curved background: as noted above, the *absence* of Hawking radiation also leads to unphysical phenomena. But there are good reasons to doubt that it is correct. In particular (following Polchinski (1995)):

1. It is possible to foliate the spacetime of a collapse-formed black hole so that curvature and energy densities on each slice remain well-behaved and far from the Planck scale (at least for black holes that are themselves large compared to the Planck mass, and up to late stages in its evaporation, of which more later).
2. The Hawking effect (if it exists) is low-energy physics, entirely describable in terms of the physics on each individual slice.
3. So the form of the cutoff imposed on our quantum-field to regularise it at short wavelengths has no effect on the low-level physics, beyond the usual effect of rescaling the parameters of the field theory (which can be absorbed by renormalisation of those parameters).
4. So it's harmless to use any cutoff we like, even the unphysical cutoff where we actually allow free-field theory to stay defined on arbitrarily short wavelengths.

In a certain sense there is even *empirical* evidence that the trans-Planckian problem is innocuous, and more generally that the arguments used to derive Hawking radiation are valid. Very close analogues of the Hawking effect occur in certain condensed-matter systems (as originally proposed by Unruh (1981)) and have recently been empirically confirmed, even though in these theories it is unambiguous that the degrees of freedom are cut off at the atomic scale and that (the analogues of) trans-Planckian modes do not exist. (See Unruh (2014) and Dardashti, Thebault, and Winsberg (2017), and references therein, for more on these analogues and their conceptual significance.)

For more on the trans-Planckian problem (and some residual worries) see Jacobson (2005, pp.46–54), Harlow (2016, pp.37–39), and references therein; however, for the moment I think we are justified in setting it aside and regarding Hawking radiation as a nigh-unavoidable consequence of any attempt to do quantum field theory in the vicinity of a black hole event horizon. Physicists tend to regard the case for Hawking radiation as further bolstered by the unity it provides to black hole thermodynamics but even without that bolstering, the case is very strong — though, of course, as good scientists we should remind ourselves that it remains purely theoretical, and that tests of quantum field theory itself in the curved-spacetime regime to date have been much less precise and numerous than in the flat-spacetime regime.

4.3. Back-reaction and evaporation

Hawking's original calculation — and all the other calculations referenced above — use quantum field theory on a fixed, non-dynamical background metric. As such, these derivations *in of themselves* do not suffice to establish that Hawking radiation is fully analogous to ordinary thermal radiation, because they imply nothing about whether a radiating black hole ultimately decreases in mass and, thus, surface area. To establish this, we need to consider the back-reaction of the radiation on the metric field, and doing so in a fully satisfactory way requires a quantum theory of gravity, which of course we lack. Furthermore, given that there is no robust local definition of gravitational energy — and, relatedly, no robust way to understand total energy as a sum of local energies — we cannot *simply* appeal to a local conservation law to conclude that radiating black holes evaporate.

Nonetheless we can give powerful arguments for that conclusion. The most direct is via appeal to Noether's theorem, applied on a sphere surrounding, and far from, the black hole: in that regime, we expect to be able to treat the hole as an approximately-isolated system in a larger region of Minkowski spacetime (see Wallace (2017b) for more on this). So the symmetries of Minkowski spacetime allow us to write a global conservation law and to argue that the sum of the ADM mass-energy of the black hole plus the total energy of the radiation outside the sphere — which is well defined, since that region is very nearly flat — should be conserved, and hence that the energy flux through the sphere ought to equal the rate of decrease of the black hole mass.

We can make this more quantitative by considering the physics on the boundary of this large sphere (here, and for the rest of this section, for simplicity I confine my attention to uncharged, non-rotating black holes). In this regime, Hawking radiation just looks like a classical outflow of radiation, with stress-energy tensor

$$T^{\mu\nu} = n^\mu n^\nu \left(A \middle/ r^2\right) \qquad (26)$$

where r is the Schwarzschild radial coordinate, n^μ is an outward-pointing null vector, and A depends on the black hole mass (and lacks a simple analytic form, due to grey-body factors). The Schwarzschild metric does not solve the Einstein field equations with this stress-energy tensor, so the assumption that the black hole does not evaporate is inconsistent with classical general relativity in a regime where we expect the latter to hold. The unique spherically-symmetric solution to the field equations for this stress-energy tensor is the *Vaidya metric* (see, e. g., Joshi, 1994), which is basically the Schwarzschild metric with a time-dependent mass term $M(t)$ ('basically' because we need to express the metric in retarded coordinates, due to the finite speed of propagation of the radiation). And the time-dependence is given by

$$\frac{\mathrm{d}M}{\mathrm{d}t} = -4\pi A, \tag{27}$$

exactly as would be predicted from a naive treatment of radiation as carrying away local mass-energy density.

To understand evaporation closer to the black hole, we need to go beyond the fully classical Einstein equation, as quantum-mechanical effects become relevant. The normal tool to investigate this is *semiclassical gravity*, in which the classical metric is coupled by the Einstein field equations to the renormalised value of the quantum expectation value of the stress-energy tensor (possibly including first-order gravitational perturbations as an additional graviton field). That is, a solution of semiclassical gravity requires both a metric g and a (Heisenberg) quantum state $|\psi\rangle$ such that

$$G[g] = 8\pi G\langle\psi\big|T\big[\widehat{\phi}\big]\big|\psi\rangle_{ren} \tag{28}$$

where $\widehat{\phi}$ schematically denotes the various quantum fields, G is the Einstein tensor associated with g, and the 'ren' subscript indicates that we need to renormalise the stress-energy tensor.

This theory can either be posited directly, on the plausible if heuristic grounds that quantum gravity 'ought' to look like this when metric fluctuations are small, or derived as the leading non-classical term in certain expansion schemes for the effective quantum field theory of gravity coupled to matter (Hartle & Horowitz, 1981; Tomboulis, 1977); either way, it is the standard tool used for exploring back-reaction (see Wald (1994, ch.5) and references therein for detailed discussion). It is difficult to calculate with; nonetheless, it has provided very strong evidence that radiating black holes do indeed radiate, and at exactly the rate predicted by the naive treatment. In particular (and without pretending to be exhaustive):

1. Candelas, Deutsch and Sciama (Candelas, 1980; Sciama et al., 1981) have calculated the stress-energy tensor for a scalar field on a Schwarzschild background near to the black hole event horizon. They find that the vacuum state of that field is strongly polarised, so as to have a very large negative stress-energy density, which diverges to negative infinity on the event horizon; this negative energy density is *exactly* cancelled out by the positive stress-energy density of the quanta in a thermal state at the Hawking temperature. It follows from their results that
   (a) the Hartle-Hawking state, in which both UP and IN modes of the field are in that thermal state, has zero net stress-energy density close to the black hole, and so solves the semiclassical equations;
   (b) any state which has any non-thermal UP mode (or any thermal UP mode at the wrong temperature) has divergent stress-energy density on the future horizon, and so fails to solve the semiclassical equations even approximately;
   (c) the Unruh state, in which the UP modes are thermally excited at the Hawking temperature but the IN modes are unexcited, has singular stress-energy density on the past horizon (which, for a collapse-formed black hole, is in any case unphysical) but only mildly nonzero stress-energy density on the future horizon, so that we would expect a self-consistent solution that is only a small perturbation of the Unruh state and the Schwarzschild solution;
   (d) In that small perturbation, the change in the area of the horizon can be calculated via the Newman-Penrose equation; the result is exactly in accord with the naive prediction from radiation flow. (This also gives insight into how the black hole's area can decrease in violation of the area theorem: the strongly polarised spacetime region close to the horizon allows a slight violation of the null energy condition.)

   Frolov and Thorne (1989) generalised these findings to rotating black holes.

2. Price, Thorne and Zurek (Zurek & Thorne, 1985; Thorne, Price, and Macdonald, 1986, ch. VIII) translated this analysis into the membrane paradigm. In that framework, the positive stress-energy associated with the atmosphere of a black hole in the Hartle-Hawking state (i. e. with both UP and IN modes thermal at the Hawking temperature) exactly cancels the negative stress-energy due to vacuum polarisation. If the IN states are unexcited, this results in a very slight depletion of the energy of the atmosphere and so a very slight negative energy flow across the stretched horizon. A precise set of conservation equations can be written at the stretched horizon that relate changes in its area to the flow of stress-energy across it; again, these reproduce exactly the naive prediction.
3. Abdolrahimi, Page, and Tzounis (2016) use numerical methods to find the metric of a radiating black hole as a perturbation of the Schwarzschild metric; they obtain a metric which far from the event horizon becomes asymptotically close to the Vaidya metric, again with the expected rate of mass decrease.

In conclusion: there are excellent reasons to think that the 'naive' treatment of radiation gets the facts exactly right: black hole radiation carries away energy and decreases the mass and surface area of the radiating black hole.

4.4. Hawking radiation and black hole thermodynamics

Hawking radiation slightly complicates the definition of 'equilibrium' for black holes, but no more so than for any other radiating thermodynamic system. Any electromagnetically-interacting body above absolute zero will radiate, so if such a system is placed alone in the vacuum, it will eventually cool to absolute zero. We can handle this in three ways:

1. Place the system in a box (of arbitrary size) filled with thermal radiation at the same temperature as the system. The radiation and the system will be in thermal equilibrium with one another and the system will itself remain at equilibrium.
2. Place the system in an empty box that is not too large. It will fill up with thermal radiation at the same temperature as the system; if it is sufficiently small, this will happen without the system's temperature changing too much.
3. Finesse the issue by ignoring radiation, on the assumption that the timescale on which it cools the object is long compared to other timescales of interest.

All these are available for black holes; the only subtlety is that the black hole's negative heat capacity means that it will be in *unstable* equilibrium with a sufficiently large thermal bath. In normal circumstances, if a small fluctuation causes the radiating system to absorb a bit of heat, its temperature rises above that of the radiation bath, so it emits the heat back again; for a black hole, that fluctuation *decreases* the temperature, so positive feedback will occur. However, if the box is sufficiently small, the decrease in temperature of the radiation bath exceeds that of the black hole and the system remains stable. Elementary calculations (Hawking, 1976) demonstrate that the total mass-energy of radiation in the box must be less than 1/4 of the black hole mass; for a solar-mass black hole, the box must be no more than $\sim 10^{12}$ parsecs across, not an especially demanding constraint.

More importantly, Hawking radiation allows black holes to be in thermal contact with one another (and with other thermodynamic systems), in just the same ways as for other self-gravitating systems. The simplest way to do this is just to put the two systems (one or both of which is a black hole) in a large box, far enough from one another that their mutual gravitational interaction can be neglected.[7] The box will fill up with radiation at a temperature intermediate between the two, and so heat will flow from the hotter body into the radiation and thence into the colder body. In particular, if the two bodies are at the same temperature, no energy will flow from one to the other: the Zeroth Law holds fully for black holes.

Alternately (and following Unruh and Wald (1982, 1983)), we can achieve thermal contact via "black hole mining": slowly lowering a box on a rope into a black hole's atmosphere, letting it fill with thermal radiation, and slowly pulling it out again. The net energy extracted from the black hole in this process is easily calculated to be

$$Q = \alpha(P+\rho)V \tag{29}$$

where V is the box's volume and α, ρ and P are respectively the redshift and the locally-measured radiation density and radiation pressure at the point where the box is removed. From the First Law of black hole thermodynamics, the change in the black hole's entropy is Q/T_H (where T_H is the hole's temperature); meanwhile, the box contains radiation at a temperature of T_H/α. Thermal radiation has an entropy density $s = (P+\rho)/T$, so the entropy increase at infinity is also Q/T_H; in other words, this process is reversible, and indeed can be reversed just by slowly lowering a box of radiation into a black hole's atmosphere until its local temperature matches that of the atmosphere, opening it, and then slowly pulling out the empty box.

If instead we try to lower the box into *another* black hole's atmosphere until we have extracted the same work as was required to lift the box in the first place, we will find that this is possible only if the second black hole is at a lower temperature than the first; if not, radiation pressure will support the box before we have extracted enough work. So — just as with the radiation spheres — this may be seen as a means of enabling heat flow from one black hole to another.

Finally, if we lower the box into the second black hole until it is exactly supported by radiation pressure — which is to say, until its temperature matches the local temperature of the atmosphere — we will find that the net work done is

$$W = (\alpha_1 - \alpha_2)(P+\rho)V \tag{30}$$

where α_1 and α_2 are the redshifts at which the box is respectively filled and emptied. If the two black holes have temperature T_1, T_2, then we must have $T_1/\alpha_1 = T_2/\alpha_2$, so that the heat Q_1 extracted from the first black hole, the heat transferred to the second black hole, and the work extracted satisfy

$$Q_2 = (T_1/T_2)Q_1; \;\; W = (1 - T_1/T_2)Q_1. \tag{31}$$

So this process is a Carnot process between the two black holes.

4.5. The generalised second law and the Casini-Bekenstein bound

At this point, we have established that stationary black holes behave *almost* exactly like thermodynamic systems. But there is a loose end left over from section 3.6: we have not yet established that the Second Law applies in full generality, nor seen how to block Geroch's and Susskind's thought-experiments which apparently allow violations of the Second Law. What would be needed to tie up this loose end would be a proof, in semiclassical gravity, of the "generalised second law": that the entropy of the black hole exterior plus the Bekenstein-Hawking entropy is non-decreasing. (As Harlow (2016, p.34) notes, "generalised second law" is a bit misleading if black hole area really *is* entropy, in which case this would just be the *ordinary* second law. On the other hand, in semiclassical gravity it is the sum of a statistical-mechanical entropy with a purely phenomenological entropy, so it does have a hybrid nature.)

The three decades following Bekenstein's original conjecture saw a substantial if rather disunified literature on various thought experiments intended to support the generalised second law. For instance, Unruh and Wald (1982) argued that the Geroch process is prevented by Hawking radiation: the global entropy maximum of a (small) box of a given energy is achieved when the box is full of thermal radiation at that energy, and that box will float, supported by the radiation pressure of the black hole atmosphere, when it is deep enough into the atmosphere that its temperature matches the local atmosphere temperature. It can readily be shown that if the box is then opened so that its contents fall into the black hole, the entropy increase of the hole equals the entropy in the box. But this argument is controversial (see, e. g., Bekenstein, 1999; Marolf & Sorkin, 2002) and in any case does not seem to address the case where a black hole is *formed* by mass with a high entropy/energy ratio. Other authors offered various more-or-less rigorous arguments for the Bekenstein bound from quantum field theory, though for a long while Bekenstein's conjecture proved difficult to make precise, and at one point was thought to rely on some ceiling on the number of distinct fundamental particles (intuitively, the more particles there are, the more states there are at a given energy). See Wall (2009) and references therein for a review of various attempts to prove the generalised second law over this period.

The last decade, however, has seen major progress in this area, largely due to increased insight into the way quantum entanglement changes with time in the black hole exterior. Building on work of Marolf, Minic, and Ross (2004), Casini (2008) was able to give a clear statement of the Bekenstein bound and then prove it fairly rigorously within quantum field theory. Similar ideas have been used by Wall (2012) to give a clear statement and fairly general proof of the generalised second law.

While no doubt there is more to learn here, and plenty of interesting foundational work to do in understanding the recent

[7] This situation is a good illustration of my comment in the Introduction about mathematical rigor. In classical general relativity it is known (Manko & Ruiz, 2001) that there is no exactly-stationary vacuum solution describing two Kerr black holes (thanks to Erik Curiel for the reference). The general approach in (most of) the black-hole-thermodynamics literature is to dismiss this sort of concern on the grounds that (a) what is needed is not exact stationarity, but approximate stationarity, i. e. negligible change of black hole orbit on the timescales relevant to the problem at hand; (b) looking for exact solutions is in any case premature given that we do not have an exact theory of black holes which incorporates radiation and back-reaction. Since the decay timescale for binary black holes scales with the fifth power of their separation, but the time taken for a radiating black hole to equilibrate with its box scales with the cube of the box size, there does not seem to be any problem of principle in constructing a setup in which the binary system can indeed be treated as approximately stationary other than thermodynamic effects.

results and their link to Bekenstein's work, there now seems to be pretty strong evidence that the generalised second law holds in semiclassical gravity in full generality, completing the case for a thermodynamic description of black holes.

5. Conclusion

Black hole thermodynamics is often described as a striking analog of ordinary thermodynamics. But if what it is to be a thermodynamic system is to obey the various laws of thermodynamics, and to interact with other thermodynamic systems in such a way that the combined system obeys those laws too, then stationary black holes are not *analogous* to thermodynamic systems: they *are* thermodynamic systems, in the fullest sense. More precisely, according to the best physics we currently have, a black hole at (or weakly perturbed from) equilibrium behaves *exactly* like a conducting, viscous fluid at (or weakly perturbed from) equilibrium, arranged in a thin shell just outside the event horizon.

An obvious question follows. In all other cases we know, there is a statistical-mechanical underpinning both to the general laws of thermodynamics, and to the specific form of the equation of state and transport coefficients of each thermodynamic system. Can we likewise construct a black hole statistical mechanics to underpin black hole thermodynamics — or are black holes fundamentally different from other thermodynamic systems at the microphysical level despite their common phenomenology? I address this topic in Part II.

Acknowledgements

I am grateful to Harvey Brown, Craig Callender, Sean Carroll, Neil Dewar, Eleanor Knox, Carina Prunkl, Katie Robertson, and Chris Timpson for various useful discussions, and in particular to Jeremy Butterfield, Erik Curiel and an anonymous referee for many helpful suggestions.

Appendix A. Dougherty and Callender on black hole thermodynamics

In a thoughtful and provocative recent paper, Dougherty and Callender (2016; henceforth DC) reach the opposite conclusion to mine: that “the analogy [between black hole thermodynamics and the ordinary kind] is not nearly as good as is commonly supposed.” They advance three arguments: that BHT “is often based on a kind of caricature of thermodynamics”; that it is ambiguous to what systems BHT is supposed to apply; that BHT is motivated by a controversial epistemic conception of entropy. Here I want to reply to these arguments.

A.1 A pale shadow of thermodynamics?

DC point out many apparent weaknesses in the details of the analogy between black hole and ordinary thermodynamics, and it is simplest to respond to them in objection-reply form.

DC: What is called the “Zeroth law” of BHT is analogous to a mere consequence of the real Zeroth Law.
Response: Fair enough (cf section 3.5). But the true Zeroth Law holds for black holes as much as for other thermodynamic systems, once Hawking radiation is allowed for (section 4.4).
DC: In ordinary thermodynamics, equilibrium systems minimise their internal energy; it's not clear whether that's even meaningful for black holes.
Response: Consider a black hole away from equilibrium. Assuming it eventually settles down to stationarity (for which, we have seen, there is strong evidence) then at late times the system will consist of outgoing gravitational radiation far from the black hole, plus a stationary black hole. To an arbitrarily good approximation we can then assign mass separately to the black hole and the radiation via Noether's theorem; gravitational radiation has positive energy, so the stationary black hole must have lower mass than its progenitor. This heuristic argument can be made precise via the Bondi mass (section 2.5): the *Bondi mass loss formula* demonstrates that a system that emits gravitational waves has decreasing mass. See, e. g., Madler and Winicour (2016) and references therein for a review of the techniques involved.
DC: There is no ‘in equilibrium with’ relation for black holes.
Response: Hawking radiation lets us define such a relation in pretty much the same way it is defined for other gravitating and/or radiating bodies (section 4.4).
DC: In ordinary thermodynamics, internal energy is distinct from total energy; in BHT, it is identified with total energy.
Response: That's an artefact of working in the black hole rest frame, which is done purely for convenience (section 3.2).
DC: If two black holes coalesce into one, the total entropy increases, even if the two black holes started off at the same temperature, ‘contrary to thermodynamics’.
Response: It's not contrary to thermodynamics. It's contrary to the thermodynamics of extensive systems, but black holes — like self-gravitating systems in general — aren't extensive (section 2.5).
DC: Substituting black hole entropy for area in thermodynamic laws makes a mess of thermodynamic relations where volume is a variable.
Response: Black hole entropy doesn't actually have the dimensions of area, unless we work in Planck units, in which case everything is dimensionless. But in any case, just because two quantities have the same units doesn't mean they can be substituted for one another in equations. (I confess I don't entirely understand DC's point here.)
DC: BHT is very non-extensive.
Response: Indeed it is, but (a) nothing in thermodynamics requires extensivity, and (b) extensivity fails in strongly self-gravitating systems for clear physical reasons (section 2.5), even before we consider black holes. (DC recognise this last point in a footnote, but claim that the subtleties of scaling in self-gravitating systems are ‘not analogous’ to those of black holes. They don't say why; examples like the radiation sphere certainly look closely analogous.)

Beyond these specific points, if DC find that BHT is a caricature of ordinary thermodynamics, it is in part because the version of BHT they are discussing is itself a caricature, pretty much restricted to the laws of BHT stated in Bardeen et al. (1973). They don't consider Christodolou and Ruffini's discussion of reversible and irreversible processes, or any of the results of the membrane paradigm, or the various results on equilibration, and more importantly, while they note the existence of Hawking radiation they don't consider its role in thermal contact, or in permitting reversible heat flow to and from black holes via Unruh-Wald mining of the thermal atmosphere.

A.2 Entropy of what?

DC point out that the event horizon is a globally defined concept, that a concept like that is not a suitable basis for BHT, and that there is no generally-agreed-upon or unproblematic alternative. They are surely right to identify this as a profoundly important question with ramifications for our understanding of Hawking radiation, and perhaps for quantum theory more generally. But it doesn't seem that relevant to black hole *thermodynamics*. After all, thermodynamics is concerned with systems at equilibrium, and is essentially silent about non-equilibrium systems except to require that they go to equilibrium. So all BHT needs is a clear understanding of the horizon for stationary black holes (and, perhaps, for holes that are mildly perturbed away from equilibrium). But pretty much all candidate definitions for the horizon agree on stationary black holes.

A.3 Entropy and empiricism

Bekenstein's original conjectures about black hole entropy made heavy use of the relation between information theory and entropy, and that link is frequently used as motivation in textbook discussions to this day. DC are critical both of information-theoretic approaches to entropy in general (they prefer a Boltzmannian conception of thermodynamics in which the information-entropy link is broken) and about its application to black holes in particular (they regard the idea of information being lost behind the event horizon as a particularly pernicious form of operationalism, given that we could just jump into the black hole ourselves at the same time that the ”lost” information falls in).

I think DC are being a little unfair here, both to Bekenstein himself (whose conjecture about black hole entropy was a consilience argument based on Christodolou, Ruffini and Hawking's results, and on various concrete thought-experiments, as much as on the general entropy-information link) and to operationalism (as shown by Hayden and Preskill (2007), in the absence of Planck-scale effects, matter thrown into a black hole will be unobservable even by an observer who jumps in after it after only time $\sim M \log M$ ($\sim 10^{-3}$ second for astrophysical-scale black holes, recall); suggestively, this is the time it takes for the stretched horizon to equilibrate in the membrane paradigm, so information thrown into a black hole is lost *in principle* after the black hole has equilibrated).

But let's stipulate that they are entirely correct. That might be a reason not to have awarded a grant to Bekenstein (or Hawking) back in the 1970s. It doesn't seem a good argument against black hole thermodynamics now, after the discovery of the Hawking effect, the membrane paradigm and the Casini-Bekenstein bound. The case for black hole thermodynamics can now rest entirely on the concrete results that have been inspired by Bekenstein's conjecture, and does not need Bekenstein's original motivation. The history of science is full of ideas whose original motivation was shaky but which nonetheless worked out, and which now stand on their own without need for that original motivation.

To be fair to DC here, in their dialectic they take themselves already to have shown that the formal analogy between black hole thermodynamics and ordinary thermodynamics is weak, so that substantial additional motivation is needed to identify entropy with black hole area. So my criticisms of this section are not really independent of my earlier points.

References

Abbott, B.e. L. S. C., & Collaboration, V. (2016). Observation of gravitational waves from a binary black hole merger. *Physical Review Letters, 116*, 061102.

Abdolrahimi, S., Page, D. N., & Tzounis, C. (2016). Ingoing Eddington-Finkelstein metric of an evaporating black hole. https://arxiv.org/abs/1607.05280.

Arnowitt, R., Deser, S., & Misner, C. (1962). The dynamics of general relativity. In L. Witten (Ed.), *Gravitation: An introduction to current research* (pp. 227–265). John Wiley and Sons. (Reprinted in General Relativity and Gravitation 40(2008) pp. 1997–2027).

Bardeen, J., Carter, B., & Hawking, S. (1973). The four laws of black hole mechanics. *Communications in Mathematical Physics, 31*, 161–170.

Bekenstein, J. D. (1973). Black holes and entropy. *Physical Review D, 7*, 2333–2346.

Bekenstein, J. D. (1999). Non-Archimedean character of quantum buoyancy and the generalized second law of thermodynamics. *Physical Review D, 60*, 124010.

Belot, G., Earman, J., & Ruetsche, L. (1999). The Hawking information loss paradox: The anatomy of a controversy. *The British Journal for the Philosophy of Science, 50*, 189–229.

Binney, J., & Merrifield, M. (1998). *Galactic Astronomy*. (Princeton: Princeton University Press).

Bondi, H. (1960). Gravitational waves in general relativity. *Nature, 186*, 535.

Brown, H. R., & Uffink, J. (2001). The origins of time-asymmetry in thermodynamics: The minus first law. *Studies in the History and Philosophy of Modern Physics, 32*, 525–538.

Callender, C. (2011). Hot and heavy matters in the foundations of statistical mechanics. *Foundations of Physics, 41*, 960–981.

Candelas, P. (1980). Vacuum polarization in Schwarzschild spacetime. *Physical Review D, 21*, 2185–2202.

Carroll, S. (2003). *Spacetime and geometry: An introduction to general relativity*. San Francisco, CA.: Addison Wesley.

Carter, B. (1979). The general theory of the mechanical, electromagnetic and thermodynamic properties of black holes. In S. Hawking, & W. Israel (Eds.), *General relativity: An Einstein centenary survey* (pp. 294–369). Cambridge: Cambridge University Press.

Carter, B. (1997). Has the black hole equilibrium problem been solved?. https://arxiv.org/abs/gr-qc/9712038.

Casini, H. (2008). Relative entropy and the Bekenstein bound. *Classical and Quantum Gravity, 25*, 205021.

Christodolou, D., & Ruffini, R. (1971). Reversible transformations of a charged black hole. *Physical Review D, 4*, 3552–3555.

Chrusciel, P., & Costa, J. (2008). On uniqueness of stationary vacuum black holes. *Astérisque, 321*, 195–265. http://arxiv.org/abs/0806.0016.

Crispino, L. C. B., Higuchi, A., & Matsas, G. E. A. (2008). The Unruh effect and its applications. *Reviews of Modern Physics, 80*, 787–838.

Curiel, E. (2014). Classical black holes are hot. https://arxiv.org/abs/1408.3691.

Damour, T. (1978). Black hole eddy currents. *Physical Review D, 18*, 3598.

Dardashti, R., Thebault, K. P., & Winsberg, E. (2017). Confirmation via analogue simulation: What dumb holes could tell us about gravity. *The British Journal for the Philosophy of Science, 68*, 55–89.

Dougherty, J., & Callender, C. (2016). Black hole thermodynamics, more than an analogy?. Forthcoming http://philsci-archive.pitt.edu/13195/.

Earman, J. (2011). The Unruh effect for philosophers. *Studies in History and Philosophy of Modern Physics, 42*, 81–97.

Frolov, V. P., & Thorne, K. S. (1989). Renormalized stress-energy tensor near the horizon of a slowly evolving, rotating black hole. *Physical Review D, 39*, 2125.

Harlow, D. (2016). Jerusalem lectures on black holes and quantum information. *Reviews of Modern Physics, 88*, 015002.

Hartle, J. B. (1973). Tidal friction in slowly rotating black holes. *Physical Review D, 8*, 1010.

Hartle, J. B. (1974). Tidal shapes and shifts on rotating black holes. *Physical Review D, 9*, 2749.

Hartle, J. B., & Hawking, S. W. (1976). Path-integral derivation of black-hole radiance. *Physical Review D, 13*, 2188.

Hartle, J. B., & Horowitz, G. T. (1981). Ground-state expectation value of the metric in the 1/N or semiclassical approximation to quantum gravity. *Physical Review D, 24*, 257–274.

Hawking, S. (1972). Black holes in general relativity. *Communications in Mathematical Physics, 25*, 152–166.

Hawking, S. (1975). Particle creation by black holes. *Communications in Mathematical Physics, 43*, 199.

Hawking, S. (1976). Black holes and thermodynamics. *Physical Review D, 13*, 191.

Hawking, S. W., & Ellis, G. F. R. (1973). *The large scale structure of space-time*. Cambridge: Cambridge University Press.

Hawking, S. W., & Hartle, J. (1972). Energy and angular momentum flow into a black hole. *Communications in Mathematical Physics, 27*, 283–290.

Hayden, P., & Preskill, J. (2007). Black holes as mirrors: Quantum information in random subsystems. *Journal of High Energy Physics, 2007*, 120.

Israel, W. (1976). Thermo-field dynamics of black holes. *Physics Letters A, 57*, 107–110.

Jackson, J. D. (1999). *Classical electrodynamics* (3rd ed.). New York: John Wiley and Sons.

Jacobson, T. (1996). *Introductory lectures on black hole thermodynamics. Lectures at the Institute for Theoretical Physics*. University of Utrecht. http://www.physics.umd.edu/grt/taj/776b/lectures.pdf.

Jacobson, T. (2005). Introduction to quantum fields in curved spacetime and the Hawking effect. In A. Gomberoff, & D. Marolf (Eds.), *Lectures on quantum gravity* (pp. 39–89). Springer. Available online at: https://arxiv.org/abs/gr-qc/0308048.

Jacobson, T., Kang, G., & Myers, R. C. (1995). Increase of black hole entropy in higher curvature gravity. *Physical Review D, 52*, 3518–3528.

Jacobson, T., & Mohd, A. (2015). Black hole entropy and Lorentz-diffeomorphism Noether charge. *Physical Review D, 92*, 124010.

Joshi, P. S. (1994). *Global aspects in gravitation and cosmology*. Oxford: Clarendon Press.

Lynden-Bell, D., & Wood, R. (1968). The gravo-thermal catastrophe in isothermal spheres and the onset of red-giant structure for stellar systems. *Monthly Notices of the Royal Astronomical Society, 138*, 495–525.

Macdonald, D. A., & Suen, W.-M. (1985). Membrane viewpoint on black holes: Dynamical electromagnetic fields near the horizon. *Physical Review D, 32*, 848.

Madler, T., & Winicour, J. (2016). Bondi-Sachs formalism. https://arxiv.org/abs/1609.01731.

Manko, V., & Ruiz, E. (2001). Exact solution of the double-Kerr equilibrium problem. *Classical and Quantum Gravity, 18*, L11–L15.

Marolf, D., Minic, D., & Ross, S. (2004). Notes on spacetime thermodynamics and the observer-dependence of entropy. *Physical Review D, 69*, 064006.

Marolf, D., & Sorkin, R. (2002). Perfect mirrors and the self-accelerating box paradox. *Physical Review D, 66*, 104004.

Misner, C. W., Thorne, K. S., & Wheeler, J. A. (1973). *Gravitation*. New York: W.H. Freeman and Company.

Parikh, M. K., & Wilczek, F. (2000). Hawking radiation as tunnelling. *Physical Review Letters, 85*, 5042–5045.

Peebles, P., & Ratra, B. (2003). The cosmological constant and dark energy. *Reviews of Modern Physics, 75*, 559–606.

Penrose, R. (1969). Gravitational collapse: The role of general relativity. Il Nuovo Cimento Numero Speziale I, 257. Reprinted. *General Relativity and Gravitation, 34*(2002), 1141–1165.

Penrose, R., & Floyd, R. (1971). Extraction of rotational energy from a black hole. *Nature; Physical Science, 229*, 177–179.

Polchinski, J. (1995). String theory and black hole complementarity. https://arxiv.org/abs/hep-th/9507094.

Robinson, S. P., & Wilczek, F. (2005). Relation between Hawking radiation and gravitational anomalies. *Physical Review Letters, 95*, 011303.

Sachs, R. (1961). Gravitational waves in general relativity. vi. the outgoing radiation condition. *Proceedings of the Royal Society of London, 264*, 309.

Sachs, R. (1962). Gravitational waves in general relativity. viii. waves in asymptotically flat space-time. *Proceedings of the Royal Society of London, 270*, 103.

Sciama, D., Candelas, P., & Deutsch, D. (1981). Quantum field theory, horizons and thermodynamics. *Advances in Physics, 30*, 327–366.

Sorkin, R. D., Wald, R. M., & Jiu, Z. Z. (1981). Entropy of self-gravitating radiation. *General Relativity and Gravitation, 13*, 1127–1146.

Susskind, L. (1995). The world as a hologram. *Journal of Mathematical Physics, 36*, 6377–6396.

Susskind, L., Thorlacius, L., & Uglum, J. (1993). The stretched horizon and black hole complementarity. *Physical Review D, 48*, 3743–3761.

Thorne, K. S., & Macdonald, D. (1982). Electrodynamics in curved spacetime: 3+1 formulation. *Monthly Notices of the Royal Astronomical Society, 198*, 339–343.

Thorne, K. S., Price, R. H., & Macdonald, D. A. (Eds.). (1986). *Black holes: The membrane paradigm*. New Haven: Yale University Press.

Tomboulis, E. (1977). 1/N expansion and renormalization in quantum gravity. *Physics Letters, 70B*, 361–364.

Unruh, W. (1976). Notes on black hole evaporation. *Physical Review D, 14*, 870.

Unruh, W. (1981). Experimental black-hole evaporation? *Physical Review Letters, 46*, 1351–1353.

Unruh, W. (2014). Has Hawking radiation been measured? *Foundations of Physics, 44*, 532–545.

Unruh, W. G., & Wald, R. M. (1982). Acceleration radiation and the generalized second law of thermodynamics. *Physical Review D, 25*, 942–958.

Unruh, W. G., & Wald, R. M. (1983). How to mine energy from a black hole. *General Relativity and Gravitation, 15*, 195–199.

Wald, R. (1993). Black hole entropy is Noether charge. *Physical Review D, 48*, 3427–3431.

Wald, R. M. (1994). *Quantum field theory in curved spacetime and black hole thermodynamics*. Chicago: University of Chicago Press.

Wald, R. M. (2001). The thermodynamics of black holes. *Living Reviews in Relativity, 4*(6). Online version at https://arxiv.org/abs/gr-qc/9912119.

Wall, A. C. (2009). Ten proofs of the generalized second law. *Journal of High Energy Physics, 2009*, 021.

Wall, A. C. (2012). A proof of the generalized second law for rapidly changing fields and arbitrary horizon slices. *Physical Review D, 85*(104049). Most recent version at https://arxiv.org/abs/1105.3445v5.

Wallace, D. (2010). Gravity, entropy, and cosmology: In search of clarity. *The British Journal for the Philosophy of Science, 61*, 513–540.

Wallace, D. (2014). Thermodynamics as control theory. *Entropy, 16*, 699–725.

Wallace, D. (2015). The quantitative content of statistical mechanics. *Studies in the History and Philosophy of modern Physics, 52*, 285–293. Originally published online under the title "What statistical mechanics actually does".

Wallace, D. (2017a). The case for black hole thermodynamics, part II: Statistical mechanics. https://arxiv.org/abs/1710.02725.

Wallace, D. (2017b). The relativity and equivalence principles for self-gravitating systems. In D. Lehmkuhl, G. Schliemann, & E. Scholz (Eds.), *Towards a theory of spacetime theories* (pp. 257–266). New York: Birkhäuser.

Wallace, D. (2017c). *Why information loss is paradoxical*. Forthcoming, https://arxiv.org/abs/1710.03783.

Weinberg, S. (1989). The cosmological constant problem. *Reviews of Modern Physics, 61*, 1–23.

Wuthrich, C. (2017). Are black holes about information? Forthcoming. In R. Dawid, R. Dardashti, & K. Thébault (Eds.), *Epistemology of fundamental Physics: Why trust a theory*. Cambridge University Press. Available at: arxiv.org/abs/1708.05631v2.

Znajek, R. (1978). The electric and magnetic conductivity of a Kerr hole. *Monthly Notices of the Royal Astronomical Society, 185*, 833.

Zurek, W., & Thorne, K. S. (1985). Statistical mechanical origin of the entropy of a rotating, charged black hole. *Physical Review Letters, 54*, 2171–2175.