Brit. J. Phil. Sci. **50** (1999), 189–229

The Hawking Information Loss Paradox: The Anatomy of a Controversy

Gordon Belot, John Earman, and Laura Ruetsche

ABSTRACT

Stephen Hawking has argued that universes containing evaporating black holes can evolve from pure initial states to mixed final ones. Such evolution is non-unitary and so contravenes fundamental quantum principles on which Hawking's analysis was based. It disables the retrodiction of the universe's initial state from its final one, and portends the time-asymmetry of quantum gravity. Small wonder that Hawking's paradox has met with considerable resistance. Here we use a simple result for C*-algebras to offer an argument for pure-to-mixed state evolution in black hole evaporation, and review responses to the Hawking paradox with respect to how effectively they rebut this argument.

1 Introduction

In August 1975 Stephen Hawking submitted to the *Physical Review* a paper entitled 'The Breakdown of Physics in Gravitational Collapse'. Published under the somewhat milder title 'The Breakdown of Predictability in Gravitational Collapse' (Hawking [1976]; see also Hawking [1982], [1998a], [1998b]), Hawking's paper did not appear until the November 1976 issue, an abnormally

long delay for *Physical Review*.[1] The paper argues that a closed system containing an evaporating black hole will evolve from a pure initial state to a mixed final state. No such evolution is unitary. Hawking took this non-unitarity to signal the breakdown of *physics* in so far as black hole evaporation is a quantum process and quantum processes are fundamentally unitary. Pure-to-mixed state evolution signals the breakdown of (one sense of) *predictability* in so far as a system in a pure state is a system for which there is some non-degenerate observable whose value may be predicted with certainty, while there is no such observable for a system in a mixed state. Pure-to-mixed state evolution signals the breakdown of (another sense of) *predictability* in so far as the mapping from initial to final states is non-invertible (*cf.* Appendix 1), and so foils the retrodiction of the initial state of a universe containing an evaporating black hole from its final state. Speaking loosely, we might say that information about the universe is lost in the course of black hole evaporation, and join the vulgar in labelling Hawking's result 'the Hawking Information Loss Paradox'. But the contention at the heart of the Hawking paradox can be stated without mention of 'information'.[2] It is the contention that in the course of black hole evaporation, pure states evolve to mixed ones. When we speak with the vulgar of 'information loss,' we mean this non-unitary evolution.[3]

The Hawking paradox has generated an extraordinary amount of interest and controversy in the physics community. Don Page's 1994 review article (Page [1994]) contains over 280 references to articles, pre-prints, and talks; each month new papers on the paradox are posted to the on-line pre-print files for High Energy Physics-Theoretical and/or General Relativity and Quantum Cosmology; a *Science* 'Research News' summary (Flam [1993]), a *Scientific American* article (Susskind [1997]), and a *New York Times* article (Johnson

[1] As reported in Page ([1994]), the title was toned down to satisfy the referee; the delay in publication may have been due to more substantive objections.

[2] There are various ways to define a measure of information and to show that the measure decreases in value in a black hole evaporation. We will not pursue this route here because it only adds a false veneer of precision and does nothing to illuminate the underlying physics.

[3] To quote Hawking ([1998b], p. 126): '[I]f the evolution is not unitary, there will be loss of information. An initial state that is a pure quantum state can evolve to a quantum state that is mixed. This process of evolution from a pure state to a mixed state is called loss of quantum coherence. It is what those who are attached to unitarity object to so violently.'A bit of hairsplitting is appropriate here. It is not loss of unitarity *per se* in black hole evaporation that is the shocker. After all, there are strong indications that even in a flat spacetime the quantum dynamics of a free scalar field cannot be unitarily implemented between arbitrarily chosen Cauchy surfaces (see Torre and Varadarajan [1998]). It is rather the pure-to-mixed transition (which implies but is not implied by loss of unitarity) that is the shocker because it is this feature that is responsible for the loss of retrodictibility, the failure of CPT-invariance, etc. The general inability to unitarily implement the dynamics for quantum fields on a curved spacetime (see Helfer [1996]) underscores the need for an algebraic approach to quantum field theory, an approach we adopt below. It also calls into question the assumption, made by many of the participants in the debate over the information loss paradox, that in non-exotic spacetimes—not involving black hole evaporation and the like—there is an S-matrix between 'in' and 'out' states defined for suitable time slices.

[1998]) alert the lay public to the paradox. What fuels the controversy is a sense that Hawking's result threatens some value the physics community holds dear.[4] Some subsets of the community evidently feel more threatened than others. For to a not inconsiderable degree, the controversy can be cast as a clash of sub-cultures in physics, with the high energy physicists typically eager (if not desperate) to avoid the paradox, while general relativists are generally more prepared to embrace it. Just as Bell's theorem invigorated investigations into the foundations of quantum mechanics by appearing to bring them within the purview of laboratory physics, the 1992 introduction of models of two-dimensional dilatonic black holes (Callan *et al*. [1992]) rejuvenated interest in the Hawking paradox, by placing processes relevantly similar to the evaporation of four-dimensional black holes in settings more amenable to tractable calculation.[5]

So reaction to the information loss paradox has identifiable social and methodological catalysts. We are less interested in investigating these than we are in examining the paradox's purported implications for the foundations of physics. Some take the CPT-invariance of quantum gravity to hang in the balance of the Hawking controversy (*cf*. Wald [1984b], reviewed in Appendix 1); we may thus add symmetry to predictability, unitarity, and the preservation of purity on our growing list of endangered valuables. The scope of the list suggests that the Hawking paradox merits the attention of philosophers of science. A taxonomy of responses to Hawking, accompanied by a sense of the commitments giving rise to them, can help us understand not only the controversy at hand, but also the nature and operation of constraints governing the construction and interpretation of physical theories. Our aim in this paper is to provide just such a taxonomy. Our plan is as follows. Section **2** sketches Hawking's result. Section **3** presents what we take to be the cleanest and most compelling argument for pure-to-mixed state transitions in black hole evaporation, an argument which furnishes principles for an analytic taxonomy. If a solution to the Hawking paradox is viable, it must deny at least one of the argument's premises; solutions may be classified with respect to which premises they deny, and evaluated with respect to the plausibility of their denial. Extant taxonomies (e.g. Preskill [1993]; Page [1980], [1994]; Stominger [1996]) tend to classify solutions with respect to their accounts of 'where

[4] Hawking's own explanation of the phenomenon is this: 'Physicists seem to have a strong emotional attachment to information. I think this comes from a desire for a feeling of permanence. They have accepted that they will die, and even that the baryons which make up their body will eventually decay. But they feel that information, at least, should be eternal' (Hawking [1998b], p. 125).

[5] Evidence for the dilaton rejuvenation may be found in *Science Citation Index*, which for the three-year period (1989–91) preceding publication of the dilaton paper lists a grand total of 15 citations to Hawking ([1976]), but lists *96* for the three-year period (1993–5) following publication. The message of these dilatonic models is mixed. Kuchař *et al*. ([1997]) develop one which is unitary, while the RST model (Russo, Susskind, and Thorlacius [1993]) reproduces all the features of Hawking's original analysis. So from these models we can draw a minor moral: the choice of boundary conditions and quantization procedure amounts to significant physics.

the information goes.' Our taxonomy enables us to discern similarities between species declared distinct by this intuitive taxonomy, and also to declare extinct those species recognized by the intuitive taxonomy which neglect to rebut Section **3**'s argument. Section **4** discusses a misunderstanding of the argument's conclusion that may have led some to overreact. In Section **5** we give an overview of different attempts to escape the information loss paradox. We reject some of these escapes out of hand, but deem others worthy of further examination. That examination is carried out in Sections **6–9**, which take up successively escapes based on thunderbolt evaporation, quantum bleaching, black hole complementarity, remnants, and baby universes. Once the overreaction to the Hawking information loss paradox has been damped down, there remain a number of unsettled issues. By our reckoning, prominent among these is the prospect of doing quantum field theory on non-globally hyperbolic spacetimes. We offer some preliminary remarks on this matter in Appendix 2. Section **10** gives our summary and conclusions.

2 The components of the paradox

The chain of reasoning constituting the information loss paradox has links which are forged by precise technical results, as well as links which are ingeniously fabricated from heuristic considerations. Our aim in this unrigorous section is simply to convey a sense of how the argument runs.[6]

What anchors the chain of reasoning is the conviction that general relativity (GR) is the correct classical theory of gravity, so that in accordance with the Penrose singularity theorem[7] black holes form in the gravitational collapse of stars. Figure 1 gives the Penrose diagram of a black hole formed in spherical gravitational collapse.

The first link is provided by the discovery of Hawking radiation: a Schwarzschild black hole is not really black, as classical GR would have it, but radiates with a blackbody spectrum at a temperature inversely proportional to the mass of the black hole.[8] Although numerous technical subtleties are needed to make the foregoing statement precise, the underpinnings are firmly established by relatively uncontroversial theory and calculational techniques developed to deal with quantum field theory (QFT) on curved spacetimes. It has become apparent that Hawking radiation is a kinematical as opposed to a dynamical effect, in the sense that Einstein's gravitational field equations do

[6] An excellent overview of the physics behind black hole evaporation is found in Wald ([1994]). There are many review articles on the Hawking information loss paradox; three which we found especially helpful are Giddings ([1993]), Preskill ([1993]), and Page ([1994]). A presentation aimed at the lay audience is Susskind ([1997]).

[7] The relevant singularity theorems are discussed in Hawking and Ellis ([1973]) and Wald ([1984a]).

[8] Hawking radiation was first proposed in Hawking ([1975]). For a review of relevant technical results, see Wald ([1994], Section 5.4).

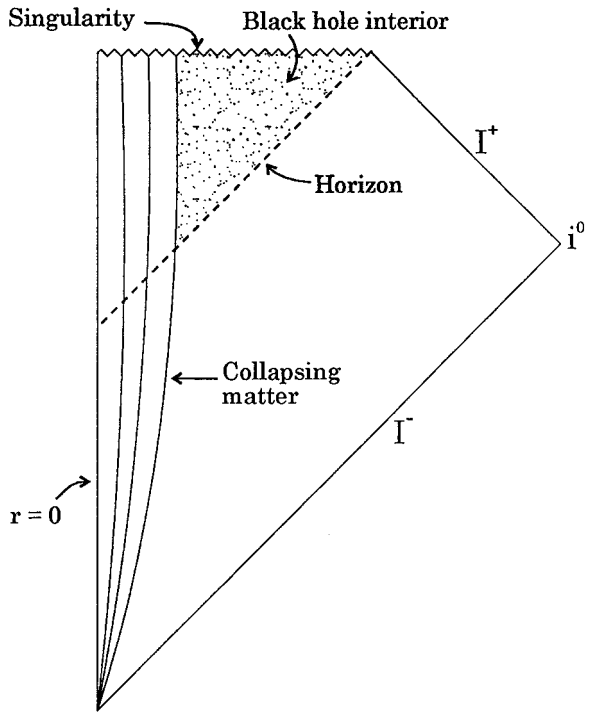

Fig. 1. Spherically symmetric black hole.

not play any role in deriving the effect—the effect holds for any quantum field which propagates in a locally Lorentz invariant manner when put on a relativistic spacetime with the appropriate event horizon structure (see Visser [1998]).

The next link is much weaker because it is based on a hybrid of GR and quantum mechanics (QM) called semi-classical quantum gravity. The idea is to take the quantum expectation value $< T_{ab} >$ of the stress-energy tensor T_{ab} of the quantum fields and then to compute the 'back reaction' on the metric g_{ab} by inserting $< T_{ab} >$ in place of T_{ab} in Einstein's field equations:

$$G_{ab} = 8\pi < T_{ab} > \tag{1}$$

(The Einstein tensor G_{ab} is defined as $R_{ab} - \frac{1}{2}Rg_{ab}$, where R_{ab} is the Ricci tensor and R is the scalar curvature.) There is very likely no self-consistent theory behind this procedure, and in any case no one has actually carried out the back reaction calculation for the Schwarzschild black hole.[9] Nevertheless, heuristic considerations strongly suggest that, at least until the Planck mass is reached, the Hawking radiation results in the *evaporation* of the black hole in the sense

[9] For reasons why, see Wald ([1994], Section 5.4).

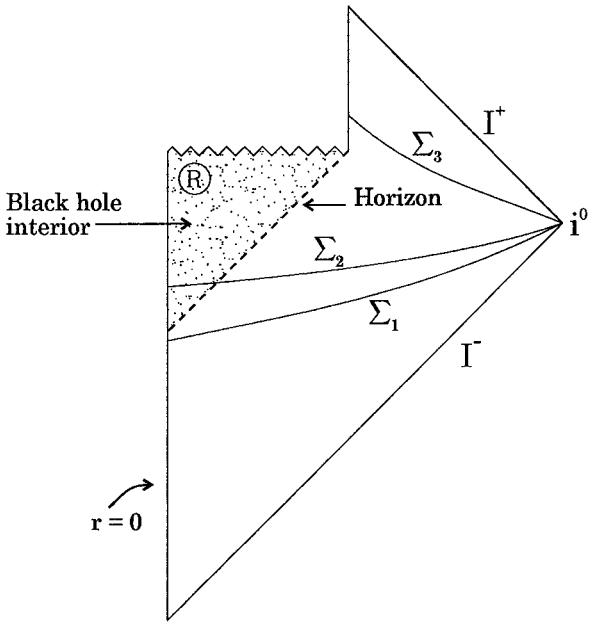

Fig. 2. Black hole evaporation.

that the process can be sensibly described by a family of Schwarzschild solutions whose mass M decreases with time at a rate proportional to $1/M^2$.

The third and fourth links are the most controversial. The third contends that there is no cut-off in the evaporation process and that the process continues until the black hole has completely evaporated. Figure 2 is a common rendering of the upshot of complete evaporation. The fourth contends that *complete* black hole evaporation leaves the universe in a mixed state. To see why, consider Figure 2's Σ_1, Σ_2, and Σ_3, spacelike hypersurfaces corresponding respectively to times before the black hole forms, after it forms but before it evaporates completely, and after it has completely evaporated. Let $\mathcal{H}_{\text{in}}$ be a space of possible states of the universe at Σ_1 and $\mathcal{H}_{\text{out}}$ be a space of possible states of the universe at Σ_3. In these terms, Hawking's heresy is that black hole evaporation inflicts a map from in-states to out-states that can be described only by a *non-unitary* superscattering matrix \$: $D(\mathcal{H}_{\text{in}}) \rightarrow D(\mathcal{H}_{\text{out}})$ (where $D(\mathcal{H})$ denotes the set of density matrices on a Hilbert space $\mathcal{H}$). To generate the heresy, consider the state of the universe at intermediate times Σ_2. This state will be an element of a tensor product space $\mathcal{H}_{\text{BH}} \otimes \mathcal{H}_{\text{ext}}$ whose components are respectively spaces of black hole interior states and black hole exterior states. The density matrix ρ_{ext} associated with the region exterior to the black hole at a time Σ_2 will describe a mixed state ($(\rho_{\text{ext}})^2 \neq \rho_{\text{ext}}$). This is because ρ_{ext} is obtained by tracing out over degrees of freedom describing the

interior of the black hole, and because the exterior and interior degrees of freedom are correlated—in particular, Hawking shows that the radiation propagating towards spatial infinity is correlated with the radiation entering the black hole. Of course, the mixed character of ρ_{ext} at Σ_2 is unexceptional. For ρ_{ext}, describing a proper subsystem of the total system, is compatible with a total state which remains pure. But consider what happens after the black hole has evaporated. Now the state ρ_{ext} just is the state of the entire universe. Mixed until the time of complete evaporation, ρ_{ext} remains mixed thereafter. So the state of the universe, originally pure (or so we assume), is now mixed. (In the next section, we provide a more rigorous argument that the post-evaporation state is mixed.) The mixed state occupied by the Hawking radiation is a function of the evaporated black hole's mass, but is largely insensitive to the detailed constitution of the matter which collapsed to form the black hole. So universes in distinct pure initial states can evolve via black hole formation and evaporation to the same mixed final state. This foils the capacity of late hour physicists to retrodict conditions in the early universe from post-evaporation data. For example, black hole evaporation should proceed in a baryon-antibaryon symmetric manner, leaving the Hawking radiation with expected baryon number 0, no matter what the baryon number of the material collapsing to form the black hole was.

The remainder of this paper evaluates various ways of avoiding Hawking's conclusion. Before turning to this task, we must address two prima facie reasons for being unimpressed by the conclusion and so for being uninterested in the ensuing debate. First, one might ask, why all the fuss? After all, according to the 'orthodox' interpretation of QM, a collapse process carries quantum systems from pure states to statistical mixtures every time they are measured. Hawking's pure-to-mixed state evolution is therefore no more remarkable than the most mundane of laboratory interactions. But to thus dismiss the Hawking paradox is to suppress the problem of quantum measurement, which is (at least) to reconcile the miracle of measurement collapse with the unitary Schrödinger equation. What's more, even if some no-collapse solution to the measurement problem (such as the modal interpretation)[10] banishes pure-to-mixed state transitions, black hole evaporation, if correctly described by Hawking, reintroduces them.

The second damper on attention to the Hawking paradox acknowledges that fundamental principles are at stake, but wonders whether it is productive to debate these principles in our present state of ignorance. Black hole evaporation is an effect that can receive a coherent and precise treatment only in a quantum theory of gravity. Since at present we have only the glimmerings of such a theory, debates about black hole evaporation are

[10] For a review of various attempts to solve the measurement problem, see Albert ([1992]).

apt to strike an impartial observer as so much thrashing around in the dark. A more optimistic point of view, one consistent with an appropriate caution regarding the soundness of the foregoing chain of reasoning, is that the thrashing may yet strike some sparks that help to illuminate the shape of the final theory. That is the attitude we will, reflectively, adopt here.

3 An argument for pure-to-mixed state transition in black hole evaporation

In this section we rehearse an argument funding our analytic taxonomy of escapes from the information loss paradox. The argument assumes that black hole evaporation is complete, that the process can be described, to some good approximation, by a spacetime of classical GR whose conformal diagram is given in Figure 2, and that the framework of local QFT is valid in this setting. In the sequel, we consider escapes that challenge all of these assumptions.

The standard approach to QFT assumes that the background spacetime is globally hyperbolic (see Wald [1994]).[11] This threatens to create a conundrum since, as we will argue below, the Hawking paradox arises in an interesting form only if black hole evaporation leads to a violation of global hyperbolicity. Fortunately, the algebraic approach to QM can be applied to (at least some) non-globally hyperbolic spacetimes to produce a QFT that is sufficient to present purposes (see Yurtsever [1994] and Appendix 2). The only extant competing approach is the sum over histories version of the path integral approach,[12] and while it has some appealing features, its foundations remain obscure.

We are usually taught to associate the (pure) states of quantum theory with vectors in a Hilbert space, to define quantum observables as self-adjoint operators acting on that space, and to imbue this formalism with empirical content by taking the expectation value of an observable O in a state $|\psi>$ to be $<\psi|O|\psi>$. The algebraic approach to QFT runs these lessons in reverse. It associates observables with elements of an abstract algebra *A*, and takes states to be positive linear functionals mapping elements of the algebra to real numbers which we understand as their expectation values in that state. One advantage of the algebraic formulation is that it enables us to finesse troubling issues concerning the unitary inequivalence of distinct and independently acceptable Hilbert space representations of QFT in curved spacetime (for

[11] For a precise definition, see Hawking and Ellis ([1973]). A necessary and sufficient condition for a spacetime to be globally hyperbolic is that it possesses a Cauchy surface, a spacelike hypersurface that meets every maximally extended causal curve exactly once.

[12] This approach has been developed by Hartle and co-workers; see Hartle ([1995]).

details, see Wald [1994], Section 4.5). Now the set of bounded operators on a Hilbert space forms a C^*-algebra,[13] and every Hilbert space state gives rise to an algebraic state in the sense of a map from elements of the algebra to real numbers. The converse holds as well: every algebraic state gives rise to a Hilbert space representation (the so-called GNS representation) in the following sense: where A is a C^*-algebra of observables and ω a state over that algebra, there exists a Hilbert space $\mathcal{H}$, a map π_ω from elements of A to bounded operators on $\mathcal{H}$, and a (cyclic) state $|\xi>$ in $\mathcal{H}$ such that $\omega(A) = <\xi|\pi_\omega(A)|\xi>$ (see Takesaki [1979]).

The argument for pure-to-mixed state transitions in black hole evaporation appeals to a basic result for C^*-algebras:[14]

Lemma. If A is a C^*-sub-algebra of a C^*-algebra B and if the restriction ω_A of a state ω on B to A is pure, then $\omega(xy) = \omega(x)\omega(y)$ for all $x \in A$ and for all $y \in B$ that commute with every element of A.

Here a state ω is said to be pure iff it cannot be written as a non-trivial convex sum: $\omega(\bullet) = \lambda_1 \omega_1(\bullet) + \lambda_2 \omega_2(\bullet)$, where $\lambda_1 + \lambda_2 = 1$, $0 < \lambda_1, \lambda_2 < 1$, and $\omega_1 \neq \omega_2$. In this setting, a pure-to-mixed state transition implies the loss of information/predictability in that a pure state, but not a mixed state, is dispersion free for some non-degenerate observable x, i.e. $[\omega(x)]^2 = \omega(x^2)$. (In the GNS construction every algebraic state, mixed as well as pure, is represented as a vector state. Mixed algebraic states are characterized by the fact that the representation is reducible.)[15]

To apply this Lemma to the case of black hole evaporation we assume there to be a global C^*-algebra B of observables associated with the entire spacetime and a global state ω defined on this algebra. If Σ is a local time slice (spacelike hypersurface) or a global time slice (spacelike hypersurface without edges),

[13] A C^*-algebra is a normed *-algebra A where the involution operation * and the norm $\| \ \|$ satisfy the condition that $\|x^*x\| = |x|^2$ for all $x \in A$.

[14] The Lemma is taken from Takesaki ([1979], p. 210).

[15] Here is an intuitive way to see the connection between a mixed state ω over a C*-algebra A and the reducibility of its GNS representation $(\mathcal{H}, \pi_\omega, |\xi>)$. If the representation is reducible, $\mathcal{H}$ will have a non-trivial invariant subspace $\mathcal{H}_1$. We can write $\mathcal{H}$ as a direct sum of $\mathcal{H}_1$ and its orthogonal complement $\mathcal{H}_2 \equiv \mathcal{H}_1^{\perp}$: $\mathcal{H} = \mathcal{H}_1 \oplus \mathcal{H}_2$. Likewise, any $O \in \pi_\omega(A)$ can be written $O = O_1 \oplus O_2$. Now let $|\psi>$ be any normed element of $\mathcal{H}$. Then $|\psi>$ can be written as $c_1|\psi_1> + c_2|\psi_2>$, $|\psi_1> \in \mathcal{H}_1$, $|\psi_2> \in \mathcal{H}_2$, $|\psi_1| = |\psi_2| = |c_1|^2 + |c_2|^2 = 1$. Calculate the expectation value of O:

$$< \psi|O|\psi > = < c_1\psi_1 + c_2\psi_2|O_1 \otimes O_2|c_1\psi_1 + c_2\psi_2 >$$
$$= |c_1|^2 < \psi_1|O|\psi_1 > + |c_2|^2 < \psi_2|O|\psi_2 >$$

since $O_1(\psi_1) \in \mathcal{H}_1$ and $O_2(\psi_2) \in \mathcal{H}_2$. This looks exactly like an expectation value for the mixed state which, in density matrix formalism, is written as $|c_1|^2 |\psi_1><\psi_1| + |c_2|^2 |\psi_2><\psi_2|$.

then we take the state at 'time' Σ to be the restriction of ω to the subalgebra of observables associated with Σ. (Here it is helpful to think of the analogy with the Heisenberg picture in Hilbert space, where there is a fixed vector state and the observables evolve. Then the only thing one could mean by the 'state at a given time' is the expectation value of the observables associated with that time. If, and only if, the evolution of the observables is unitary can the switch be made to the Schrödinger picture.) Following Wald ([1994]), we take the algebra of observables associated with Σ to be $A(\mathrm{Int}(D(\Sigma))$, i.e. the algebra associated with the open region $\mathrm{Int}(D(\Sigma))$, the interior of the domain of dependence $D(\Sigma)$ of Σ.[16]

Now take A in the Lemma to be $A(\mathrm{Int}(D(\Sigma_3)))$, where Σ_3 is a 'time' after the black hole has completely evaporated (cf. Figure 2). Consider any open region R in the black hole interior. Since R and $\mathrm{Int}(D(\Sigma_3))$ are relatively spacelike, any y in $A(R)$ should commute with any x in $A(\mathrm{Int}(D(\Sigma_3)))$. Let us call this the *commutation condition*. But one expects that for some such x and y, $\omega(xy) \neq \omega(x)\omega(y)$. Let us call this the *correlation condition*. Hawking's analysis furnishes a reason to expect the correlation condition to hold. According to Hawking, radiation emerging from the black hole will be correlated with radiation falling into it. He offers a heuristic physical model of these formal correlations. Particle–antiparticle pairs are created near the black hole's horizon; negative energy particles fall into the black hole, decreasing its mass, while positive energy particles emerge from the vicinity of the horizon as Hawking radiation. On this model, ingoing and outgoing particles share a common causal past—a region of overlap between their backward light-cones—in which a correlation-establishing common cause mechanism operates. Thus for some x in $A(R)$ and y in $A(\mathrm{Int}(D(\Sigma_3)))$—where R and $\mathrm{Int}(D(\Sigma_3))$ are spacelike separated regions sharing a common causal past—$\omega(xy) \neq \omega(x)\omega(y)$.

[16] The domain of dependence $D(\Sigma)$ of Σ is defined as the union of the future $D^+(\Sigma)$ and past $D^-(\Sigma)$ domains of dependence. The former is defined as the set of all spacetime points p such that any past endless causal curve through p meets Σ. The latter is defined analogously. There are two motivations for the choice of $A(\mathrm{Int}(D(\Sigma)))$ as the algebra associated with Σ. The first is that it does no harm for purposes of showing that the state at 'time' Σ is mixed; for if the restriction of ω to this algebra is mixed, then so is the restriction to any subalgebra, and surely the state at 'time' Σ will be the restriction of the global state to some subalgebra of $A(\mathrm{Int}(D(\Sigma)))$. The second and more positive motivation stems from the following property of primitive causality which should be satisfied in a reasonable QFT. Let Σ be a (local or global) time slice and let $\tilde{\Sigma}$ be any 'thin sandwich' about Σ. Then $A(\mathrm{Int}(\tilde{\Sigma}))$ is irreducible with respect to $A(\mathrm{Int}(D(\Sigma)))$ in that any bounded element of the latter that commutes with all the elements of the former is a multiple of the identity. This means that any bounded element of $A(\mathrm{Int}(D(\Sigma)))$ is a function of the elements $A(\mathrm{Int}(\tilde{\Sigma}))$, and that if the two algebras are von Neumann algebras then they are identical. (A bounded operator is said to be a function of a set of operators if it commutes with every bounded operator that commutes all of the operators of the given set. The only operators that commute with an irreducible set of operators are multiples of the identity, and since any bounded operator commutes with the identity, any bounded operator is a function of an irreducible set of operators. Von Neumann algebras are C^*-algebras that are identical with their double commutants and are weakly closed under taking bounded functions.)

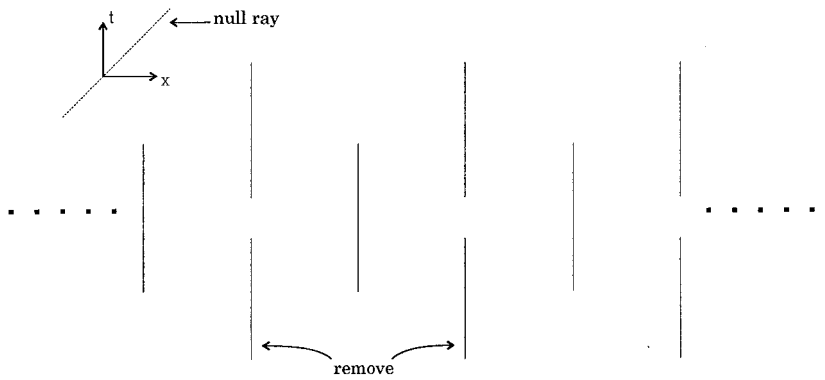

Fig. 3. Mutilated two-dimensional Minkowski spacetime.

While we may be conditioned by the literature on EPR-type experiments to such distant correlations as somewhat exotic, in quantum field theory, they are endemic.[17] Now the correlation and commutation conditions enable us to invoke the Lemma to conclude that the state at 'time' Σ_3—that is, the restriction of ω to $A(\mathrm{Int}(D(\Sigma_3)))$—is mixed. Those who prefer Hilbert space talk to algebraic talk can rephrase the result in terms of the Heisenberg picture; the Schrödinger picture, of course, is unavailable since unitarity fails.

At this juncture a cautionary remark is in order. To complete the argument that black hole evaporation involves a pure-to-mixed state transition one needs the further proviso that the state at a time prior to evaporation (e.g. Σ_1 in Figure 2) is pure. This pre-evaporation purity proviso is not entirely innocuous; Hawking's paradox is pointed only in non-globally hyperbolic spacetimes, and the application of the Lemma to some especially ill-behaved non-globally hyperbolic spacetimes implies that the state associated with *any* (connected) spacelike Σ is mixed. Figure 3's two-dimensional Minkowski spacetime, surgically mutilated by the removal of timelike strips, is an example of such a spacetime. Of course, the Lemma does not *force* the state associated with Σ_1 of Figure 2 to be mixed. There nothing guarantees that the commutation condition holds, because there is no open region which is relatively spacelike with respect to $\mathrm{Int}(D(\Sigma_1))$. However, the Lemma provides only a sufficient condition for the state associated with a time slice to be mixed. States not revealed to be mixed by the Lemma may be mixed anyway. Nevertheless, it would be surprising if a pure state cannot be assigned to Σ_1.[18] In any case, the manoeuvre of denying that

[17] An example of what we have in mind is the standard Minkowski vacuum state which exhibits EPR-type correlations. Clifton *et al*. ([1997]) have shown that any QFT that admits states with these correlations admits a dense set of such states. The result refers to a Hilbert space representation rather than to a C^*-algebra. Presumably it can be rephrased in algebraic terms.

[18] But see Myers ([1997]), who argues that black holes cannot form from pure states!

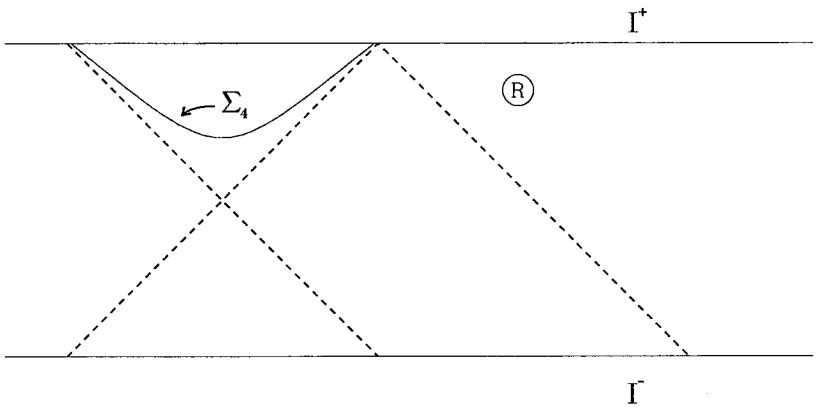

Fig. 4. De Sitter-like spacetime.

the state of the universe was pure at the outset averts one species of non-unitary evolution at the cost of rendering purity unattainable to begin with—a blow to quantum foundations arguably more violent than Hawking's! So even in the absence of a positive argument for it, we will proceed as though the pre-evaporation purity proviso is secure.[19]

4 Misunderstandings

There is an implication of the preceding two sections that we want to emphasize, for its neglect can provoke serious misunderstanding and distress. The pure-to-mixed state transition in black hole evaporation implies a failure of unitarity. But this failure of unitarity need not imply that the local laws of field propagation have been altered. Indeed, the implicit assumption behind the foregoing analysis is that the standard field laws apply locally, and consequently that unitarity can be maintained locally (in the sense of a sufficiently small globally hyperbolic neighbourhood).

The pure-to-mixed state transition and the consequent failure of unitarity at issue in black hole evaporation derives not from the breakdown of locally unitary field laws but from the global structure of the spacetime of Figure 2. This spacetime displays the exotic features of a black hole region and the failure of global hyperbolicity. But even in spacetimes free of black holes and globally hyperbolic, mixed states can be induced by the choice of time slices that are not Cauchy surfaces. Consider Figure 4's two-dimensional spacetime with manifold R^2 and the de Sitter metric $ds^2 = -dt^2 + \cosh^2(t)dx^2$. Apply the

[19] A positive argument is provided by Wald ([1980b], [1984b], [1994], Section 7.3). But it is based on the analogy with a static spacetime, whereas black hole evaporation involves a spacetime that is not even stationary. (Recall that a spactime is stationary just in case it admits a timelike Killing vector field. A spacetime is static just in case it is stationary and the Killing field is hypersurface orthogonal.)

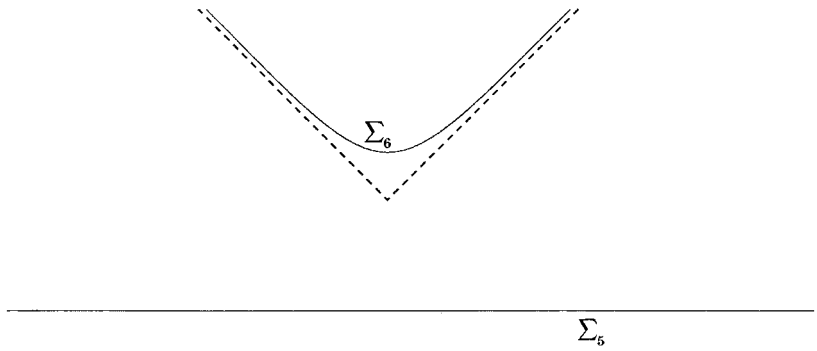

Fig. 5. Two-dimensional Minkowski spacetime.

Lemma of Section **3** to the non-Cauchy slice Σ_4 . That the region R is relatively spacelike with respect to Int(D(Σ_4)) secures the commutation condition, and that the common causal past of Int(D(Σ_4)) and R is non-empty, allowing a common cause to establish correlations between the observables on Int(D(Σ_4)) and R, secures the correlation condition. So the state associated with a non-Cauchy slice of this non-exotic spacetime is mixed.

The Lemma does not force a mixed state upon the hyperboloid slice Σ_6 of Minkowski spacetime (see Figure 5) because there is no open region R relatively spacelike with respect to Int(D(Σ_6)). Still, one might expect the failure of unitarity for the evolution up to the time Σ_6 of a massless scalar field. For one can describe an initial state on the earlier Cauchy slice Σ_5 which would have the field propagate off to spatial infinity without registering on Σ_6, suggesting a loss of probability and ergo unitarity. Wald ([1994], Section 7.3) makes this expectation rigorous. He argues, for a massless scalar field whose field equation is conformally invariant, that a pure-to-mixed state transition can take place from time Σ_5 to Σ_6. Minkowski spacetime can be conformally embedded in the Einstein static spacetime (*cf.* Figure 6). In the embedding spacetime the complement of (the embedded image of) D(Σ_6) (the shaded region of Figure 6) does contain relatively spacelike open regions (e.g. region R of Figure 6) so that the Lemma now applies. States which are pure for Minkowski spacetime and the Einstein static spacetime are mixed when restricted to Int(D(Σ_6)).[20]

The moral of such examples is that the restriction of a pure global state to the algebra of observables associated with a non-Cauchy region is liable to be mixed.[21] As no post black hole evaporation time-slice is a Cauchy slice, it

[20] We are indebted to Robert Wald for clarifying this point for us.
[21] Only liable—recall the argument reviewed in the last section that the state associated with non-Cauchy, pre-evaporation slice Σ_1 is pure!

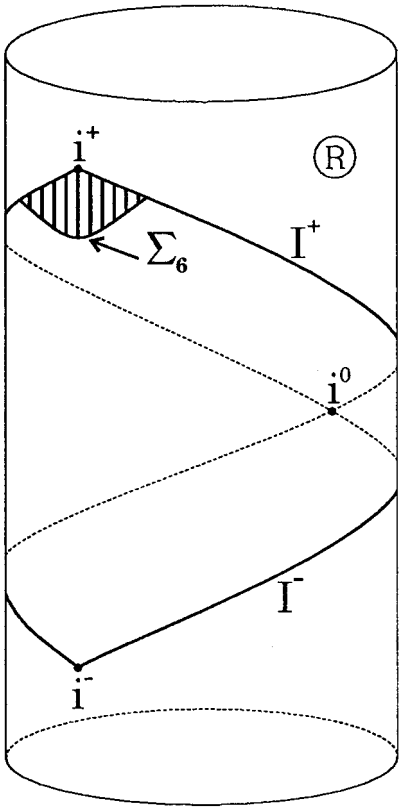

Fig. 6. Conformal embedding of Minkowski spacetime into the static Einstein universe.

should not be surprising that post-evaporation states are mixed. In light of this moral, staunch resistance to the conclusion that black hole evaporation involves a pure-to-mixed state transition is somewhat puzzling, at least if the resistors accept Figure 2 as an accurate rendering of the evaporation process. Such resistance is less interesting if it consists simply in the failure to absorb or appreciate the moral. It is much more interesting if it arises from the following claims: first, that to the extent that the reasoning behind the information loss paradox can be trusted, it must be taken to indicate that macrolevel pure-to-mixed state transitions affects microphysics; and second, that this will lead to highly unpalatable if not disastrous consequences. Interesting resistors include Banks and Susskind ([1984]), who argue that a generalization of the Schrödinger equation to allow for pure-to-mixed state transitions leads to a violation of locality and/or energy-momentum conservation, and Srednicki ([1993]), who argues that a pure-to-mixed state transition means that energy conservation and Lorentz invariance cannot both hold. Unruh and Wald ([1995]) have responded that such pathologies can be confined to 'Planckian states,' so that

ordinary laboratory physics can continue unimpeded.[22] The interesting resistance's first claim is also dubious. It is based on the notions that if GR and QM are married by quantizing the metric and that if appreciable quantum fluctuations occur at the Planck scale, then the formation and evaporation of Planck radius black holes issues pure-to-mixed state transitions on this scale (see Hawking [1982]). The first conjunct of the antecedent will be denied by those who seek to consummate the marriage of QM and GR without quantizing the metric field. But even if both conjuncts of the antecedent are accepted, the consequent can be denied since it is equally plausible that fluctuations of the metric on the Planck scale mean that the concepts of Lorentzian geometry, event horizons, Hawking radiation, etc. all break down. In other words, the Hawking information loss paradox may provide a guide to what a quantum theory of gravity must say about the macrolevel without giving much guidance about what the theory says at the microlevel. So the interesting resistance can itself be resisted.

Whether information loss is an unexceptional consequence of global spacetime structure or a harbinger of the disintegration of tractable microphysics is an important question meriting detailed examination. But we will not attempt to adjudicate it here. Instead, we will concentrate on attempts to *escape* the information loss paradox. Should some of these escape attempts seem desperate, bear in mind that they may yet be justified if they are needed to maintain locality and/or energy-momentum conservation. To be frank, however, some escape routes strike us as so bizarre or so nearly incoherent that we would rather tolerate violations of locality or energy conservation than follow them.

5 Evading the Hawking paradox: overview

What is surprising about the literature on the Hawking information loss paradox is not so much the volume of the reaction as the variety of evasive manoeuvres undertaken. There are different ways to classify the escapes, and we claim no particular virtue for ours beyond its capacity to highlight crucial distinctions not always evident in the literature. While we make no claim of completeness, we think that our classification scheme does capture the majority of the major escape routes.

Escape 1. The information loss paradox is intolerable. The culprit is Einstein's GR, which must be rejected in favour of a theory of gravity allowing stars to undergo gravitational collapse to compact objects without forming the event horizon structure constitutive of a black hole. Moffat has championed such an

[22] The question of locality and/or conservation of energy violations in dynamical schemes that generate pure-to-mixed transitions is also an interesting issue to pursue from the point of view of so-called collapse solutions to the measurement problem in QM.

alternative theory of gravity (see Moffat [1993]; Cornish and Moffat [1994]). While this line of enquiry is not to be dismissed out of hand, we will not pursue it here. To keep the discussion focused, we will simply assume that GR is the correct classical theory of gravity.

Escape 2. GR is the correct classical theory of gravity. Black holes do form and do evaporate. However, the evaporation does not take place in the way indicated in Figure 2 but rather as in Figure 7. With this 'thunderbolt evaporation' there is no loss of global hyperbolicity. As a result, the spacetime can be foliated by a family of Cauchy surfaces in such a way that there is no time at which the Lemma forces the state of the universe to be mixed. Only 'bad' choices of time slices give rise to the appearance of information loss. We will discuss the pros and cons of this escape in Section **6**.

Escape 3. GR is the correct classical theory of gravity. Black holes do form and do evaporate as shown in Figure 2. But the argument given in Section **3** is unsound because the correlation premise is false. At *no time* are black hole exterior observables correlated to black hole interior observables. Colloquially, information carried by the matter falling into the black hole is 'bleached out' at the event horizon. To our knowledge, no one currently advocates this position (though Page ([1980]) approves of something akin to it). Still, it is widely taken up in the literature as a useful straw man. Section **7** rehearses the bayonet practice that has been done on this straw man.

Escape 4. GR is the correct classical theory of gravity. Black holes do form and

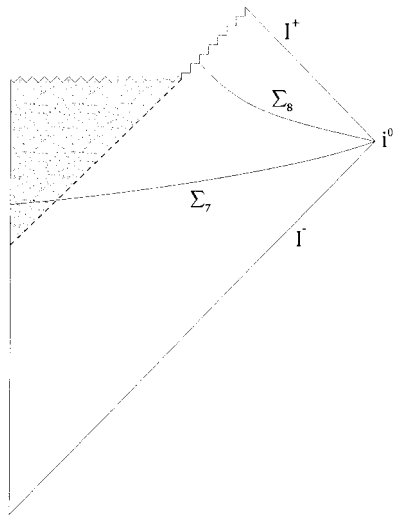

Fig. 7. Thunderbolt evaporation.

evaporate as shown in Figure 2. There is no quantum bleaching. But information is not lost because it 'leaks out' of the black hole in such a way that an external observer in the asymptotically flat region could in principle recover, by sufficiently precise measurements, the information apparently lost in black hole evaporation (Page [1980], [1994]). So cast, the escape enjoys a surface plausibility. However, it is not clear that it denies either the correlation or commutation condition of Section **3**, and so not clear how it avoids Section **3**'s conclusion that the post-evaporation state is mixed. So let us try to make it more precise. Recall that we are accepting the framework of local QFT for the purpose of considering the Hawking paradox from the standpoint of extant physics. (As we shall see, string theorists reject the framework.) In that framework's terms, take 'complete information is available at the post-evaporation time Σ_3' to mean that the set S of observables associated with time Σ_3 is such that any observable pertaining to the spacetime is a function of observables in S. The idea behind this gloss on completeness is that anything we want to know about the spacetime we can learn by measuring the elements of S. Consider an observable O associated with the black hole interior. What does it mean to say O is a function of observables in S? A generally accepted necessary condition is this: O is a function of observables in S only if O commutes with any O′ that commutes with every element of S. Now let R be an open set in the interior of the black hole. If Section **3**'s commutation condition held, every observable associated with R would commute with every element of S. According to the condition above, for R observables to be functions of observables in S, R observables, observables associated with the black hole interior, would all have to commute with one another. But it is not the case that these observables form a commuting set! So to maintain that the post-evaporation region harbours complete information (in this sense), one must deny Section **3**'s commutation condition.

We do not think that such a denial is what is intended by those who hold that information leaks out of the black hole. Nor have advocates of leakage provided any plausible mechanism for violating the commutation condition (as might be provided, for example, by string theory; *cf*. Section **8**). Some authors (ones who admit their responses to the Hawking paradox are only partial) have provided models of Hawking radiation that, by considering factors such as the mode dependence of the barrier penetration of the infalling matter, or backreaction effects, or stimulated emission, imply that the radiation is not exactly thermal (e.g. Berkenstein [1993] and Schiffer [1993]). But none of these models implies that the post evaporation exterior state is pure. Some might be taken to gesture toward the rejection of the commutation condition—for example, the model of Danielson and Schiffer ([1993]) requires non-local effects—but none makes the rejection explicit, and none takes the rejection to be its key point. Thus, in so far as Escape 4 is not confused or is not based on an

equivocation on the concept of 'information', we suspect it is a disguised way of endorsing some other escape route.

Escape 5. GR is the correct classical theory of gravity. Black holes form and they evaporate as depicted in Figure 2. There is no bleaching. But complete evaporation does not leave the universe in a mixed state, because the mixture-inducing correlations between particles interior and exterior to the black hole before it evaporates are mimicked by correlations between early and late hour Hawking radiation found in the exterior region after evaporation. Wald likens this reconstitution of purity to the unitary cooling of a material body by the emission of photons: 'The photons emitted at early times are correlated with the atoms which emitted them, and these correlations are then gradually transferred to the photons emitted at later times' ([1994], p. 184). While Wald thinks the purity reconstitution manoeuvre the most 'plausible' escape route, he considers it fatally flawed. For it posits correlations between early hour (i.e., when the causal structure classical GR attributes the spacetime should be heeded as approximately correct) Hawking radiation and interior states of the black hole. And Wald charges any mechanism for establishing such correlations with 'gross violation of causality' ([1994], p. 184). For at times early in the evaporation process the Hawking radiation is outside the forward light cone of the black hole interior states to which any such mechanism would correlate it.

Perhaps advocates of purity reconstitution would embrace gross causality violations in order to deny Section **3**'s commutation condition, and thereby disable our argument for pure-to-mixed state transitions. But they have another option. While rejecting bleaching to allow correlations between interior and exterior observables at times *up to* evaporation, purity reconstitutors can deny correlations between *post*-evaporation exterior observables and observables pertaining to the black hole interior, thereby undoing the correlation condition of the mixture-imposing Lemma. Which option purity reconstitutors adopt is an issue muddied by the fact that the most prominent camp of purity reconstitutors are Black Hole Complementarians, who supplement their physical models with an account of right discourse according to which it is meaningless to speak at all of global states and of correlations between the interior and exterior regions of the black hole, and so meaningless to speak in the Lemma's terms. But Black Hole Complementarians also have a positive story to tell. They postulate a 'stretched horizon' lying outside of the black hole event horizon. As an object falls through the stretched horizon, external observers see it thermalized. No information is lost because the information contained in the object is preserved in the outgoing radiation from the stretched horizon. The connection between this positive account and the proscription on certain ways of speaking is supposed to be this: because there is no bleaching, infalling observers do not

agree with the description just given of what happens to them as they cross the stretched horizon. A new form of complementarity, Black Hole Complementarity, prevents this disagreement from developing into a full-blown contradiction. We discuss Black Hole Complementarity and some of its radical consequences in Section **8**.

Escape 6. Black holes do form and do evaporate. But the evaporation stops when the Planck scale is reached, leaving Planckian remnants which code up the missing information. Most of the discussion of this idea has focused on the plausibility of the notion that a Planck-sized remnant could support enough quantum states to retain the missing information. Our qualms are quite different. When one tries to implement this escape route in terms of the spacetime structure of classical GR, the attempt seems incoherent; for either it does not avoid the Lemma, or else it manages to avoid the Lemma by denying the existence of black holes. We discuss remnants in Section **9**.

Escape 7. Black holes do form and do evaporate as shown in Figure 2. But the information is not lost because it is encoded in histories and not just in states associated with instantaneous time slices (Hartle [1998]). On one level we find this no more a solution than saying that information is not lost because it resides in the spacetime. Yes, in the four-dimensional atemporal sense, correlations do exist between regions R and $D(\Sigma_3)$ in Figure 2, and in that atemporal sense no information is lost. But it is precisely because these correlations exist that the state at time Σ_3 is mixed. Still, read in a more sympathetic light, Hartle's sum-over-histories approach to QM and QFT is consonant with the morals we want to draw. First, once the nature of the 'information loss' that occurs in Figure 2 is properly understood, there is no reason to be vexed. And second, the real problem raised by the black hole evaporation of Figure 2 is not the information loss *per se* but the problem of how to do QFT on non-globally hyperbolic spacetimes. The sum-over-histories approach championed by Hartle is one answer; we describe another in Appendix 2.

Escape 8. Classical GR implies 'no hair' theorems to the effect that after formation a black hole settles down to a state characterized entirely by its mass, electric charge, and angular momentum, all other features (such as multi-pole moments) having been radiated away (see Wald [1984a]). QM holds out the possibility that black holes are characterized by additional conserved quantities ('quantum hair'). If numerous enough, these quantities might plug the information loss (see Ellis *et al*. [1991] and Coleman *et al*. [1992]). The idea is that the internal state of the black hole could be uniquely fixed by the values of these quantities, and if the values are ascertainable by means of measurements made in the exterior region of the black hole, then even after a complete evaporation of the black hole, the total information is still in principle

available. In keeping with our discussion of Escape 4, we take the quantum hair proposal to escape our Lemma by rejecting the commutation condition. Notice that a hair-based rejection of the commutation condition need not be *ad hoc*, for the failure of this condition could be explained by the non-local effects associated with the quantum hair (see Preskill's ([1991]) discussion for a way to grow quantum hair from the Aharonov–Bohm effect).

Our worry about this escape route has to do not with non-locality but with a more fundamental matter. The notion that the internal black hole state can be read off the values of the quantum hair observables is problematic. Unless quantum bleaching occurs, observables interior and exterior to the black hole will be correlated, and the state of the entire system entangled. That is, the black hole interior will occupy no pure state. We can illustrate this by supposing that there is an eigenbasis $|\,i>_{BH}$ for $\mathcal{H}_{BH}$ such that the total state for $\mathcal{H}_{BH} \otimes \mathcal{H}_{ext}$ evolves into the form $|i>_{BH} \otimes |\varphi,\ v_i>_{ext}$. Here v_i values for the quantum hair observables such that if $i \neq j$ then $v_i \neq v_j$, so that $|i>_{BH}$ may be inferred from v_i. In a general entangled state $\Sigma_i\, |\,i>_{BH} \otimes |\varphi,\, v_i>_{ext}$ the quantum hair observables will not have definite values at all, much less values that tell us anything about the interior state. Indeed, in a general entangled state, there won't be a pure interior state! Of course, on the orthodox account of measurement, measuring quantum hair values will collapse the entangled composite state to a hair eigenstate associated with the values obtained. Post-collapse, measured values do suffice to determine the state of the black hole interior. But from the point of view of protecting unitarity, we are hardly better off: measurement collapse itself is non-unitary!

We now turn to a more detailed examination of some of these escapes.

6 Thunderbolt evaporation

Since black hole evaporation is a quantum gravitational effect, classical GR cannot on its own resolve the information loss debate. Nevertheless, there are results in GR which have been taken to indicate that a black hole cannot evaporate completely without leaving a naked singularity. Wald ([1984b]) gives one such result. This theorem is compatible with Figure 2 and, thus, with the conclusion of the Hawking information loss paradox. But an inspection of the theorem's premises shows that it is also compatible with the thunderbolt evaporation pictured in Figure 7 (see Earman [1995], Ch. 3). It might seem that this makes no real difference, since the argument used to prove that the state at time Σ_3 in Figure 2 is mixed applies equally well to the post evaporation time Σ_8 of Figure 7. But the crucial difference lies in the fact that the spacetime of Figure 7 is globally hyperbolic and, thus, can be partitioned by a one-parameter family of Cauchy slices. The algebra of observables associated with the domain of dependence of any such Cauchy slice is the global algebra; the

'restricted' state for the slice is pure if the global state is. There is no time (Cauchy slice) at which a state initially pure is mixed, and so no information lost. The moral is that in thunderbolt evaporation, the illusion of information loss results from the 'bad' choice of a non-Cauchy surface Σ_8, exactly analogous to the apparent loss of information in the 'bad' choice of Σ_4 in Figure 4 or Σ_6 in Figures 5 and 6.

There remains the seeming paradox that relative to any Cauchy slice in Figure 7 black hole evaporation lies to the future. That is, for 'good' choices of times, black hole evaporation never occurs! One might complain that the foregoing account of what happens as a consequence of black hole evaporation simply avoids the problem by postponing it indefinitely. The complaint assumes that 'as a consequence of' has a temporal meaning, so that an account of what happens as a consequence of black hole evaporation must be an account of what happens after a black hole has evaporated. Thunderbolt evaporation makes available a Cauchy slicing underwriting all the temporal relations anyone would need but undermining questions about what happens after evaporation. But this is no reason to get agitated—we can still sensibly describe the entire spacetime in which the black hole forms and evaporates. The complaint about the Cauchy slicing rings hollow because Newtonian intuitions about space and time do not suffice to describe what happens in general relativistic spacetimes.

In classical GR, Roger Penrose's cosmic censorship hypothesis conjectures that naked singularities do not develop for generic initial conditions.[23] The strongest form of censorship requires that the spacetime be globally hyperbolic. So one can extend the cosmic censorship hypothesis to black hole evaporation by conjecturing that (in so far as the process can be described by a classical GR spacetime) evaporation does not generically result in a violation of global hyperbolicity. If this conjecture were true, it would effectively plug Hawking's information leak.

Hawking and Stewart ([1993]) investigated four toy (1 + 1)-dimensional models of black hole evaporation. Numerical simulations were taken to indicate that the two models with high symmetry give naked singularities, whereas the two more general models give thunderbolts. However, it is far from clear whether the results are due to a breakdown of the numerical simulation and/or the semi-classical approximation, or whether they show that a thunderbolt singularity cuts off future development and prevents the formation of a naked singularity.

7 Quantum bleaching and quantum xeroxing

Bleaching and xeroxing resolutions to the Hawking paradox are typically

[23] For a discussion of the various versions of cosmic censorship, see Earman ([1995]).

posed only long enough to be dismissed (see Preskill [1993] and Giddings [1993]), and posed in terms of the intuitive but amorphous desideratum of 'information preservation'. Neither resolution survives reformulation in terms of more precise notions.

Information falling into a black hole destined to evaporate is not lost, the informal xeroxing idea goes, so long as a copy of it is left outside the event horizon. To avert information loss, we need only suppose Σ_1 to Σ_2 evolution to proceed as follows:

$$|A>_{in} \rightarrow |A>_{int} \otimes |A>_{ext} \tag{2}$$

Xeroxing ensures that the eventual evaporation of the black hole leaves the universe at post-evaporation time Σ_3 in the pure, and fully informative, state $|A>_{out}$. Since it factorizes, the r.h.s. of (2) establishes no correlations between interior and exterior observables, and thereby disarms the correlation condition on Section **3**'s argument that the Σ_3 state is mixed.

Designed to preserve information, the xeroxing model fails to uphold the gold standard of unitarity. For the model would require an initial state $|B>_{in}$ to evolve as follows:

$$|B>_{in} \rightarrow |B>_{int} \otimes |B>_{ext} \tag{3}$$

If xeroxing is linear, (2) and (3) imply

$$|A>_{in} + |B>_{in} \rightarrow |A>_{int} \otimes |A>_{ext} + |B>_{int} \otimes |B>_{ext} \tag{4}$$

But if xeroxing is xeroxing, the initial state $|A>_{in} + |B>_{in}$ must evolve into the state $(|A>_{int} + |B>_{int}) \otimes (|A>_{ext} + |B>_{ext})$ which differs from the r.h.s. of (4) by the presence of cross terms. The xeroxing proposal cannot rescue unitarity from Hawking's onslaught.

But (reverting once again to loose talk) xeroxing is not the only way to keep information outside the black hole. For it might be that it never falls in to begin with! This is the idea behind the quantum bleaching proposal. To obtain pure post-evaporation states, bleaching posits a unitary global state evolution which averts correlations between observables in the black hole interior and observables elsewhere. Such an evolution must take a pure Σ_1 state to a Σ_2 state which factorizes into a pure interior state and a pure exterior state:

$$|\psi>_{in} \rightarrow |\varphi>_{int} \otimes |\chi>_{ext} \tag{5}$$

If the Σ_2 state fails to factorize, its Schmidt decomposition[24] will have at least two terms, and the interior and exterior Schmidt bases will be perfectly correlated. To avert correlations for arbitrary initial states, require each

[24] Recall that the Schmidt decomposition of a two component tensor product state $|\Psi>$ takes the form $|\Psi> = \Sigma c_i |a_i> \otimes |b_i>$, where c_i are complex numbers whose square moduli sum to one, and the Schmidt bases $\{|a_i>\}$, $\{|b_i>\}$ are orthonormal sets of vectors on the first and second factor spaces respectively.

$|\psi'>_{in} \neq |\psi>_{in}$ to evolve into a Σ_2 state which factorizes:

$$|\psi'>_{in} \rightarrow |\varphi'>_{int} \otimes |\chi'>_{ext} \quad (6)$$

Correlation-averting evolution must be linear to assuage worries about unitarity. So (5) and (6) imply

$$|\psi>_{in} + |\psi'>_{in} \rightarrow |\varphi>_{int} \otimes |\chi>_{ext} + |\varphi'>_{int} \otimes |\chi'>_{ext} \quad (7)$$

But (7) averts correlation only if either (i) $|\chi>_{ext} = |\chi'>_{ext}$ or (ii) $|\varphi>_{int} = |\varphi'>_{int}$. Option (i) amounts to the claim that black hole exteriors occupy the same (pure!) state no matter how those black holes form, and so illustrates that preservation of purity is not on its own sufficient to resolve worries about 'information loss'. Only option (ii) is viable: correlations are unitarily averted and information preserved only if a black hole forms in the same state ($|\varphi>_{int}$) no matter what the state of the matter collapsing to form it ($|\psi>_{in}$) was. In homely terms, however you stain infalling matter, it gets bleached white at the event horizon.

The standard objection to quantum bleaching is that it can be enforced by no plausible mechanism. Quantum bleaching, if it occurs, occurs in a regime where the validity of classical gravity is unquestioned; classically, the event horizon is not a boundary marked by any sort of physical process or structure. As Strominger puts it, '[T]he horizon is a smooth place at which all curvatures are subPlanckian. There are no guards stationed there which strip intruders of all information' (Strominger [1996], p. 736). Black hole complementarians (see Section **8**) would insist that the failure of the horizon to appear special to infalling observers is perfectly compatible with its appearing cataclysmic to distant observers. But this compatibility is purchased at the cost of denying the validity of the global perspective in whose terms the bleaching model is developed, and so cannot consistently underwrite a mechanism enforcing that model. The impasse is generic: any bleaching mechanism would be (if not incompatible at least) in serious tension with basic commitments bleaching is intended to protect.

8 Black hole complementarity

Black Hole Complementarians (e.g. Susskind, Thorlacius, and Uglum [1993]) accept Figure 2 as an approximately correct representation of the evaporation process and reject quantum bleaching and xeroxing while maintaining as an axiom the 'S-matrix ansatz' that 'a unitary S-matrix describes the evolution from infalling matter to outgoing Hawkinglike radiation' (p. 3743). That this blatantly contradicts the standard analysis of black hole evaporation does not trouble the Complementarians, for their stated purpose is to develop a physical theory distinct from standard, semi-classical QFT on curved spacetime (Susskind and Thorlacius [1994], pp. 972–3); Susskind, Thorlacius, and Uglum [1993], p. 3757). Their second axiom is that a semi-classical field theory describes physics outside the 'stretched horizon' of the black hole to

good approximation. This stretched horizon is a membrane w hose area is roughly one Planck length greater than that of the black hole horizon[25] and which provides the boundary conditions for the field theory. The idea is simple enough: an observer restricted to a certain region of spacetime (a Rindler wedge of Minkowski spacetime, say, or the exterior region of a black hole) can postulate the existence of degrees of freedom on the boundary of this region which encode the physics occurring outside the region. Thus the degrees of freedom of the stretched horizon encode the physics happening in the black hole. The Complementarians' third axiom is that the space of states available to a black hole of mass M, understood in terms of the degrees of freedom associated with its stretched horizon, has the dimension of its Bekenstein entropy. It follows from this that an accurate rendering of the stretched horizon outruns the resources afforded by QFT. For the degrees of freedom field theory must attribute the stretched horizon so that Hawking radiation comes out with the desired spectrum do not give rise to the desired value for the Bekenstein entropy (*cf*. Banks [1995]). Thus the stretched horizon is to be treated not in field theoretic but in string theoretic terms. Black Hole Complementarity as a project in creative physics is the pursuit of a string-theoretic, stretched membrane description of black holes. Not everyone agrees that a string-theoretic/stretched membrane approach to black hole evaporation will rescue the information Hawking believes lost. Banks ([1995]) endorses the picture, but reckons that its eventual execution will reveal that some information will be lost to the external observer. So we should distinguish between the Complementarians' strings-and-membranes strategy for doing black hole physics and the S-matrix ansatz they hope by this strategy to preserve.

Now the payoff of the Complementarians' approach is this: as an object falls through the stretched horizon, an exterior observer will see it thermalized. The information it carries reappears later in the Hawking radiation. That is, while early Hawking radiation will be correlated with degrees of freedom of the stretched horizon (herein lies the Complementarians' denial of bleaching), as

> more time elapses [. . .] the stretched horizon emits more quanta. The previous correlations between the stretched horizon and the radiation field are now replaced by correlations between the early part of the radiation and the newly-emitted quanta. In other words, the features of the exact radiation state that allows S_E [a measure of the entanglement of interior and exterior states] to return to zero are long-time correlations spread over the entire time occupied by the outgoing flux of energy (Susskind, Thorlacius, and Uglum [1993], p. 3759).

Recalling Wald's discussion of the cooling material body, the Complementarians here appear to purchase the purity of the post evaporation state by denying the correlation condition of Section **3**.

[25] Definitions of the stretched horizon differ; see Susskind *et al*. ([1993], Section IIIA).

But rejecting the correlation condition of Section **3**'s argument may not be the only way Complementarians might evade its conclusion that the post evaporation state is mixed. One might expect them to find the rejection of the commutation condition (*cf*. Escape 4) congenial, since they tend to be string theorists. Strings are non-local objects, so it would not be surprising if string observables associated with relatively spacelike regions failed to commute. Indeed it has been shown that for interacting strings the commutator for two string fields does not vanish outside the string light cone (Lowe, Susskind, and Uglum [1994]). 't Hooft ([1997]) remarks that the S-matrix ansatz and the commutation condition are incompatible. He takes the S-matrix ansatz to imply the availability of complete information in the post-evaporation region, in the sense that the operators associated with that region afford an irreducible representation of the field theory. But if some operator associated with the black hole interior commuted with every operator associated with the post-evaporation exterior, the exterior algebra would be reducible. So 't Hooft denies the commutation condition, and attributes this violation of the causality conditions of QFT to the excessive energy carried by some of the Hawking radiation. 'Due to these energies, the space-time metric is distorted, and the light cone will not always stay in position' ([1997], p. 5).

Black Hole Complementarians speak sometimes as though they deny the correlation condition of Section **3**'s argument, other times as though they deny the commutation condition. At still other times, they take a third, and most disorienting, tack. To the physics discussed thus far the Complementarians append an analysis of meaning by whose lights the premises of that argument are nonsense (so, presumably, too are the denials of those premises, such as the denial of the commutation condition 't Hooft issues above!). To see why they think they need to employ such a theory of meaning, compare the exterior observer's description of the infalling observer's fate (she gets thermalized at the stretched horizon, then rebroadcast as Hawking-like radiation) with the infalling observer's own. Complementarians accede to the standard wisdom that since the event horizon is a global object unmarked by any local physical boundary, there is no way an observer falling through it would experience anything untoward there. So as far as the observer is concerned, she and her information make it through the event horizon intact—they are not vaporized by the membrane Complementarians postulate. From this it follows that 'the reality of the membrane cannot be an invariant which all observers agree upon' (Susskind, Thorlacius, and Uglum [1993], p. 3760). Even worse, the actuality of an event cannot be an invariant:

> Black hole complementarity and its realization in string theory imply profound changes in our current views of matter and spacetime. These concepts further erode the classical realism of the Newtonian picture of the universe. They entail a degree of relativity and observer dependence of

> reality. The special theory of relativity destroyed the invariant meaning of simultaneity [. . .] What was left intact was the invariant event, occurring in a well-defined spacetime location [. . .] Now, however, even that can no longer be relied upon (Susskind [1994], p. 6611).

For one observer matter evaporates at the stretched horizon, for another it survives to be destroyed by tidal forces as it falls toward the singularity. Inconsistency—not to mention the demise of the very idea of a spacetime description, and the tradition of modern physics funded by that idea—threatens.

Complementarians avert disaster by insisting that these apparently contradictory commitments can be simultaneously asserted only from a standpoint which is 'unphysical':

> The assumption of a state [. . .] which simultaneously describes both the interior and the exterior of a black hole seems suspiciously unphysical. Such a state can describe correlations which have no operational meaning, since an observer who passes behind the event horizon can never communicate the result of any experiment performed inside the black hole to an observer outside the black hole. The above description of the state lying in the tensor product space $\mathcal{H}_{\mathrm{BH}} \otimes \mathcal{H}_{\mathrm{out}}$ can only be made use of by a 'superobserver' outside our universe. As long as we do not postulate such observers, we see no logical contradiction in assuming that a distant observer sees all infalling information returned in Hawking-like radiation [. . .] (Susskind, Thorlacius, and Uglum [1993], p. 3744).

Rejecting global states, Complementarians reject the very terms of Section **3**'s analysis.

We should resist a *kneejerk* temptation to grumble here about a retreat to a long-ago discredited operationalist/verificationist philosophy of science. For at crisis points in science, operationalism has served to motivate new and fruitful approaches, the most famous example being Einstein's insistence on giving distant simultaneity an operational meaning. Still, there is scope for *considered* grumbling. Notice that once Einstein had constructed the special and general theories of relativity, he quickly jettisoned the operationalist philosophy. The philosophy did for him the progressive work of motivating theory construction rather than the dubious work it does for the Complementarians of averting counterexamples to the theory of meaning they append to their work in theory construction. Another ground for considered grumbling is that to the extent that Complementarians' ban on talk of global states is well taken, it applies not only to the evaporation pictured in Figure 2 but also to the standard black hole configuration pictured in Figure 1. But if one is not allowed to talk about global states for generic black hole configurations, it is not clear to us how to pose the issue of black hole evaporation, since it is not clear how to apply the theorems that underwrite Hawking radiation. It seems that for Complementarians it is problematic to speak globally about black hole *evaporation* because it makes no sense to speak globally about black holes in the first place.

Complementarians feel driven to the idea that the exterior and the free-falling observers give *complementary* descriptions of black hole evaporation, that no report issued outside an event horizon can be contradicted by any report issued inside, because no two such reports can ever be simultaneously entertained. But can't they? Complementarians worry about scenarios such as the following (*cf*. Susskind and Thorlacius [1994]). An EPR pair of particles with anti-correlated spins is created. Particle 1 falls into the black hole, where its x-spin is measured. The black hole, obedient to the Complementarians' axioms, emits radiation from which particle 1's spin state can be determined. An external observer uses this radiation to determine the z-spin of particle 1, then falls into the black hole where she receives a message telling her of the outcome of the interior x-spin measurement of particle 1. *Pace* the doctrine of Black Hole Complementarity, she is then in a position to simultaneously entertain assertions made from putatively complementary perspectives—her own past assertion, issued in the black hole's exterior, about the z-spin of particle 1, and the assertion, issued in the black hole interior, about its x-spin. Susskind and Thorlacius are sufficiently troubled by this development to argue that if our intrepid observer is to gather all the relevant data before hitting the singularity, the message from the x-spin measurement must be sent via quanta of energy that exceed the Planck scale. Complementarians conjecture that what holds here holds in general: that any experiment simultaneously recognizing the standpoint of putatively 'complementary' observers will require Planckian physics to decode, and so lies beyond present ken. This may be, but it amounts to protecting the (already distinctly flawed) program of Black Hole Complementarity by cloaking it in the unfathomed mysteries of Planckian physics. All told, Black Hole Complementarity, taken as a theory of meaning which defuses the argument for information loss by denying the meaningfulness of the terms in which that argument is cast, is far less satisfying than escapes from the Hawking paradox which offer or sketch specific mechanisms of avoidance, schemes such as quantum bleaching, thunderbolt evaporation, or the formation of remnants (to be examined in the following section).

But Black Hole Complementarity as creative physics is a program for realizing a string-theoretic mechanism for averting information loss. (Notice that this program moreover furnishes a reason for rejecting the terms of Section **3**'s analysis: those terms are field theoretic, and thus inadequate to the needs of black hole physics. So in some sense the most philosophical plank of the Complementarian platform is also the most otiose.) And we would hardly wish to discourage Black Hole Complementarians from pursuing a consistent, string-theoretic treatment of black hole evaporation. Indeed string theorists have expressed optimism that such a treatment of black hole evaporation will reveal that unitarity is not lost (see Horowitz [1997]). Perhaps this will turn out

to be correct. But we cannot see why this solution, if indeed it is a solution, need appeal to Black Hole Complementarity *qua* theory of meaning.

9 Remnants

Since the semi-classical picture of black hole evaporation cannot be trusted when the black hole has shrunk down to Planck scale, it could be that quantum gravity effects will prevent further evaporation, leaving a stable or a long lived remnant. Such Planck mass remnants would effectively constitute new species of elementary particles. Since a star collapsing into a black hole can carry an arbitrary amount of information, there would presumably have to be a continuous infinity of such species, raising worries about unbounded pair production (Preskill [1993]). Debates also rage in the literature about the feasibility of forming Planck-sized objects which are able to code up enough information to rebut Hawking. Skirting these debates here, we will instead concentrate on what seems to us a more fundamental difficulty, which we will formulate in the form of a dilemma: Either remnants are remnants—that is, of black holes—in which case they do not provide for a satisfying resolution of the Hawking paradox, or they are not remnants—at least, not of black holes—in which case they can do nothing to address the problem of *black hole* evaporation. We also examine a startling version of the remnants proposal which takes each remnant to be a universe in its own right.

Incomplete thunderbolt evaporation might leave a remnant black hole. But not even complete thunderbolt evaporation leads to information loss, or so we have argued. A remnant of this form is not needed to resolve the information loss paradox.

A way of picturing remnants which does not involve a residual black hole is given in Figure 8 (copied from Giddings [1995]). The spacetime in question still has the event horizon structure constitutive of a black hole, so while the remnant (the ??? of Figure 8) is not a remnant black hole, it is remnant of a black hole, and so confronts the dilemma's first horn. In this situation one can proclaim as loudly as one wants that information is stored in the remnant. Be that as it may, observables in the algebra associated with post-evaporation slice Σ_{10} of Figure 8(a) (stable remnant) or Σ_{11} of Figure 8(b) (long-lived remnant) ought to commute with observables associated with the black hole interior. And the presence of the ???, which indicates that the singularity of classical GTR has been replaced by a Planckian object, does not alter our standard expectation that post-evaporation observables will be correlated with interior observables. We can thus apply the argument of Section **3** to conclude that there is information loss in the sense that at time Σ_{10} or Σ_{11} the state is mixed. Of course, what the proponents of remnants may be implicitly claiming is that the descriptive apparatus used in the Lemma is incomplete and must be

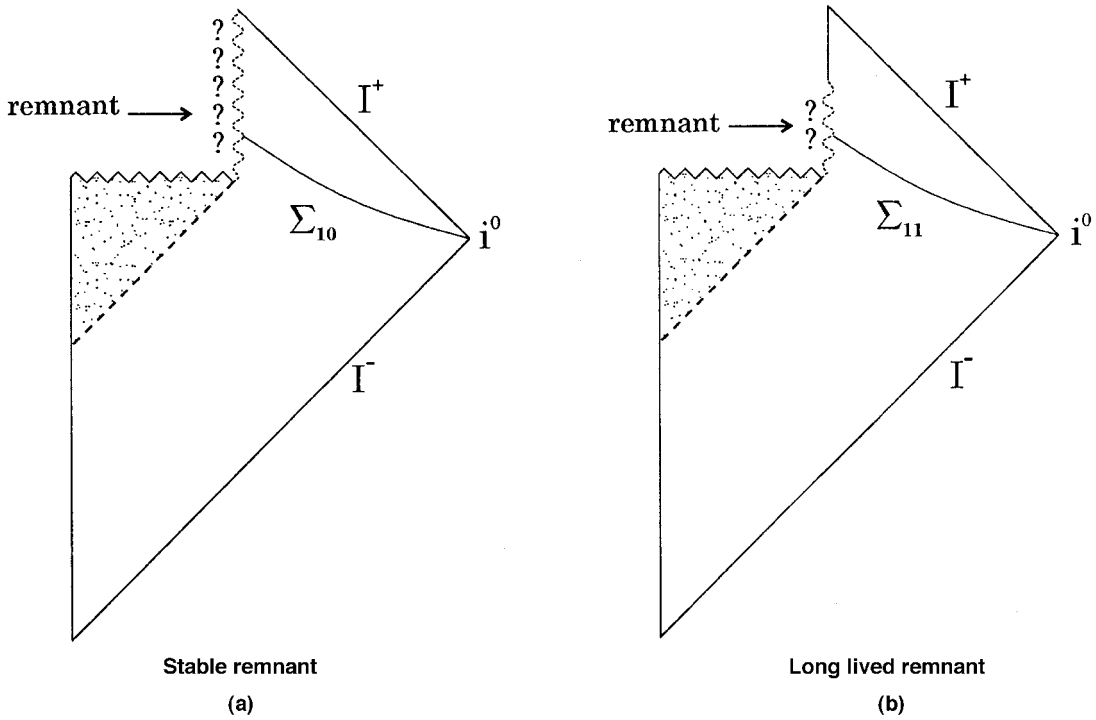

Fig. 8. (a) Stable remnant. (b) Long-lived remnant.

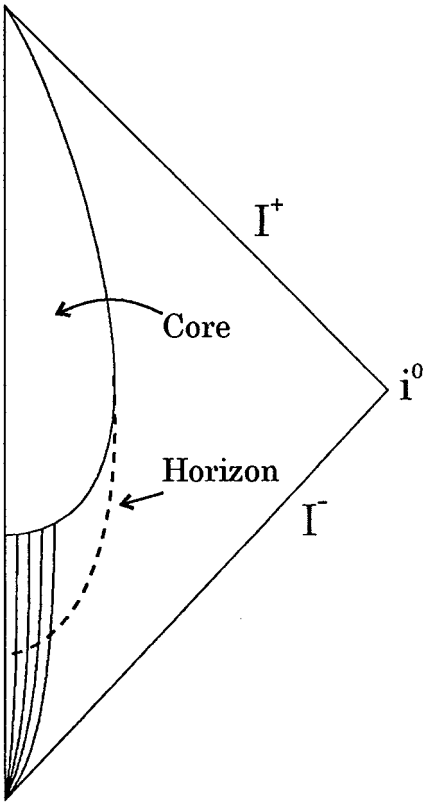

Fig. 9. Gidding's evaporation scenario.

supplemented. The ??? could, for instance, simply be a placeholder for a set of boundary conditions which, when imposed at the singularity, induce a one-to-one and invertible map from initial to final states of the universe. (Parikh and Wilczek ([1998]) show how to do this for a charged black hole evaporating into a naked timelike singularity.) But until remnant enthusiasts produce the new physics that incorporates the boundary conditions in a natural way, the present proposal 'solves' the information loss paradox only by inserting the missing information by hand, and 'remnant' is just a name that does nothing to justify the sleight of hand.

Giddings ([1992]) has sketched a third remnant scenario. While this scenario supposes that black hole evaporation halts before reaching the Planck scale, the spacetime diagram presented (Giddings [1992], Figure 3) applies equally well whatever the scale of the remnant.[26] Figure 9's remnant is meant to carry all the

[26] This diagram is hard to interpret since the usual conventions of Penrose diagrams seem to be violated.

information left out of the outgoing radiation. When the remnant emerges from the horizon, this information becomes causally accessible to observers in the asymptotically flat region. It is claimed that this resolves the information loss paradox. We are not so sure. One disturbing feature of this proposal is that the core expands at a superluminal rate. Worse, Figure 9 incorporates no singularity of gravitational collapse. (Thus, the Penrose singularity theorem must be avoided either by a violation of Einstein's field equations or the energy conditions.)[27]Because the spacetime lacks a genuine black hole, the surface labelled 'horizon' in Figure 9 must be an apparent horizon, an object locally delineated, rather than a true event horizon, a global feature marking the boundary between what can and cannot be seen from future null infinity I^+. And because the relevant event horizon structure is missing, the theorems that underwrite Hawking radiation do not apply. In short, the labels of 'black hole'' and 'black hole evaporation' strike us as misnomers when applied to Figure 9. Its remnant is not a remnant black hole It therefore seems to us that the Giddings ([1992]) version of the remnant scenario confronts the second horn of our dilemma. It is less a solution to the information loss paradox than a sweeping denial of the problem. But perhaps that is his point.

The final form of the remnant idea we will consider involves the use of 'baby universes.' Figure 10 (which is copied from Polchinski and Strominger ([1994]), Figure 1) is supposed to illustrate the formation of a black hole as the result of an infalling matter pulse. But contrary to the standard scenario, the black hole interior does not terminate in a curvature singularity. Nor does the evaporation cease, leaving a remnant in the manner of Figures 8 or 9. Rather the evaporation eventuates in a remnant that is a universe in its own right—a baby universe that branches off from the main universe. To an observer inhabiting the asymptotically flat post-evaporation region of the parent universe, there seems to be a loss of information. But this is an illusion. The information is not lost since it is contained in the baby universe that is now causally inaccessible to our observer.

In terms of the argument of Section **3**, the Lemma can be applied to show that the states at times Σ_{13} or Σ_{14} are mixed. But the baby universe rejoinder would be that the transition from a pure state at time Σ_{12} to the mixed states at the later times Σ_{13} and Σ_{14} does not signal information loss since the state at time Σ_{15}, which is the disjoint union of Σ_{13} and Σ_{14}, is pure. Because Section **3**'s Lemma offers only a necessary condition for purity, this last step remains to be justified. And there are independent reasons to worry whether the transition from time Σ_{12} to time Σ_{15} can be unitary. For classical general relativistic spacetimes a change in spatial topology—such as is involved in the branching off of a baby universe—implies the failure of global hyperbolicity (Geroch

[27] The standard singularity theorems in GR require that the stress-energy tensor T_{ab} satisfy various conditions, such as the weak and dominant energy conditions (see Wald [1984a]). These conditions can be violated by quantum fields.

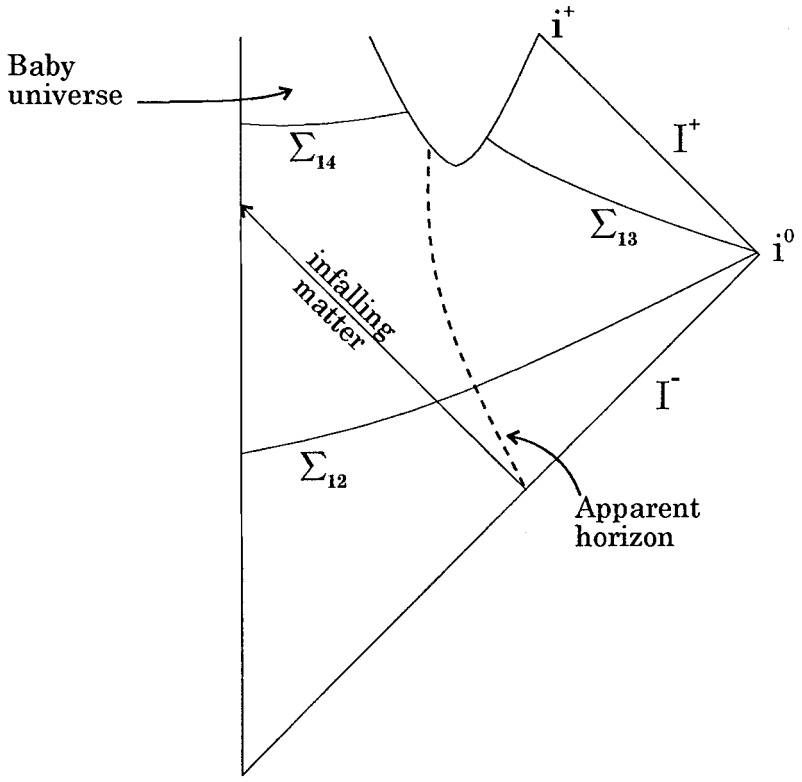

Fig. 10. Baby universe remnant.

[1967]), and unitarity for quantum fields on non-globally hyperbolic spacetimes is not to be expected in general.

In light of these qualms, it is at least initially reassuring that Baby Universe advocates furnish a positive case that baby universes render black hole evaporation unitary. Polchinski and Strominger ([1994]) show that for certain initial states of the parent universe—eigenstates of the baby-universe creation operators (!)—the map from initial parent universe state to final exterior + baby universe state is given by a unitary S matrix. Now a parent universe whose initial state is a superposition of these eigenstates would not enjoy unitary evolution. But, Strominger reminds us, the baby universe creation operator is an operator for a spacetime causally isolated from the exterior spacetime in which physicists work. This operator should therefore commute with every operator associated with our spacetime, and so divide the Hilbert space for measurements we can perform into superselection sectors out of which dynamics for the exterior spacetime cannot carry us. Within each sector, the value for the baby universe creation operators 'are effectively constants and the black hole formation/evaporation process is

effectively unitary' (Strominger [1996], p. 755). We can think of two ways to promote effective unitarity to unitarity full stop. One is the by-now familiar self-stultifying manoeuvre of invoking measurement collapse. If our initial measurements collapse the universe family into the corresponding eigenstate of the baby universe creation operator, subsequent evolution of the universe family will be unitary. But the collapse *won't* be, so such an attempt to upgrade effective unitarity falls short of the goal of preserving the fundamental unitarity of quantum processes. The second attempt invokes a runaway operationalism to collapse effective and genuine unitarity into one and the same notion. Neither is compelling.

10 Conclusion

A complete understanding of black hole evaporation will undoubtedly require new physics. But in the absence of that new physics, we are constrained to confront the problem in terms set by existing physical theories. Thus, our conclusion is necessarily qualified by assumptions that will eventually have to be revised or abandoned. But within these qualifications it is also sweeping. Assuming that black hole evaporation can be described by a spacetime of classical GR, that the evaporation is not of the thunderbolt type, and that the quantum aspects of the problem can be described by QFT on the resulting spacetime, then *none* of the escape routes discussed in the literature provides a plausible way to avoid the conclusion that the post-evaporation state is mixed. So *if* the system begins in a pure state, a pure-to-mixed state transition accompanies black hole evaporation. In this sense, information is lost, as Hawking originally maintained. To be sure, the assumptions are far from safe. The spacetime aspects of black hole evaporation can, at best, be described approximately by classical GR; and the quantum aspects can, at best, be described approximately by QFT on a spacetime of classical GR. The difficulty here is not merely that the conclusion has to be given an 'approximately' qualification, but that we cannot know how good the approximation is, or even what 'approximately' means, until we know how to combine QM and GR in one theory. Absent such knowledge, we cannot infer that the failure of unitarity in the approximate description carries over to the correct description.

Even if it does carry over, the pure-to-mixed transition hardly seems to merit what can only be described as the measures of desperation some would adopt to avoid it—at least not if the failure of unitarity is tied to the large scale structure of spacetime and does not affect the local propagation of the quantum fields. Arguments to the contrary are less than convincing. And even if arguments for the local breakdown of unitarity succeed, there may be ways to contain the ensuing violations of locality and/or energy conservation so that laboratory physics is not affected.

The Hawking paradox bears implications for determinism and predictability. Excluding thunderbolt evaporation, black hole evaporation involves a breakdown

of global hyperbolicity. In classical GR, the failure of global hyperbolicity entails a loss of predictability since (intuitively speaking) the laws of GR do not place any restrictions on what nasty things can emerge from naked singularities. But even assuming the naked singularities are classically benign (e.g. they do not ooze any classical green slime), a new element of unpredictability emerges as soon as one tries to do quantum theory on a spacetime featuring a singularity left naked by black hole evaporation. This breakdown of predictability, the original official message of the Hawking paradox, seems to us to remain unshaken. If the breakdown of predictability carries with it a breakdown of CPT and T invariance (as it seems to, *cf*. Appendix 1), it could harbour implications for the problem of the direction of time which remain to be developed (see Liu [1993]). The potential of such implications to steer us toward the correct quantum theory of gravity likewise remains to be developed. This potential notwithstanding, physicists should not pursue the issue of pure-to-mixed state transitions in black hole evaporation to the exclusion of other issues in the vicinity, prominent among which is the issue of how to do QFT on non-globally hyperbolic spacetimes. The heterogeneity of commitments, presuppositions and tolerances in the physics community, a heterogeneity laid bare by the foregoing anatomy of the Hawking information loss controversy, promises that that community is unlikely to be paralysed by an over-restrictive focus.

Acknowledgements

We are obliged to Rob Clifton, Fred Kronz, John Norton, Carlo Rovelli, Steve Weinstein, and an anonymous referee for helpful comments on earlier drafts of this paper.

Department of Philosophy
Princeton University
Princeton, NJ 08544
USA
gord@princeton.edu

Department of History and Philosophy of Science
University of Pittsburgh
Pittsburgh, PA 15260
USA
jearman@pop.pitt.edu

Department of Philosophy
University of Pittsburgh
Pittsburgh, PA 15260
USA
ruetsche+@pitt.edu

Appendix 1: Black hole evaporation and the CPT invariance of quantum gravity

Suppose that black holes form and that they evaporate. Following Hawking, grant that their evaporation carries the universe from a pure state to a mixed one. Absent a theory of quantum gravity, we cannot describe the detailed dynamics of black hole evaporation. But we do not need quantum gravity to build a quantum field theory for regions where quantum gravitational effects are negligible, regions lying in the asymptotic past of an evaporating black hole, or regions lying in its asymptotic future. Let $\mathcal{H}_{in}$ and $\mathcal{H}_{out}$ be Hilbert spaces of states on asymptotically past and asymptotically future regions respectively. Where $D(\mathcal{H}_{in})$ and $D(\mathcal{H}_{out})$ are sets of density matrices on $\mathcal{H}_{in}$ and $\mathcal{H}_{out}$, introduce a 'super scattering' matrix \$: $D(\mathcal{H}_{in}) \rightarrow D(\mathcal{H}_{out})$ whose action we understand as follows: where the density matrix W_{in} gives the initial state of a universe containing an evaporating black hole, $\$W_{in} = W_{out}$ gives its final state. We can formulate in terms of \$ what we have agreed we *can* say about the dynamics of black hole evaporation: \$ takes pure states to mixed ones.

There is widespread agreement that \$ threatens some values physicists hold dear, less agreement about which. Wald takes 'the evolution of a pure state to a density matrix [to] imply a breakdown of retrodictability' ([1984b], p. 168); Page ([1994]) describes pure-to-mixed state evolutions which leave retrodictability (in some form) intact. But Page's manoeuvre may not secure other valuables against the Hawking paradox. Wald has argued that any theory of quantum gravity adequate to black hole evaporation will not be CPT invariant. This appendix reviews Wald's argument, and looks at proposals which might seem to moderate the force of its conclusion.

Where S is the set of solutions to whatever may be the fundamental dynamical equations of quantum gravity, and f(-) maps a solution to its CPT reverse, CPT invariance requires

$$\text{(QG INV) } s \in S \text{ iff } f(s) \in S$$

To extract the consequences of quantum gravity's CPT invariance for the scattering theory encapsulated by \$, introduce $\theta: D(\mathcal{H}_{in}) \rightarrow D(\mathcal{H}_{out})$, a map from states in $\mathcal{H}_{in}$ to their CPT reverses in $\mathcal{H}_{out}$. θ is invertible; its inverse maps states in $\mathcal{H}_{out}$ to their CPT reverses in $\mathcal{H}_{in}$. If quantum gravity is CPT invariant, our scattering theory will satisfy

$$\text{(\$ INV) } (W, W') \text{ is a possible universe for \$ iff } (\theta^{-1}W', \theta W) \text{ is.}$$

To say that the ordered pair (X,Y) is a *possible universe* for \$ is to say that X and Y are physically possible initial and final states, respectively, and that $Y = \$X$. For now, take every state $W \in D(\mathcal{H}_{in})$ to be a possible initial state. If (\$ INV) holds, \$ has an inverse $\$^{-1} = \theta^{-1}\$\theta^{-1}$. What Wald shows is that $\$^{-1}$

cannot exist. $ is not invertible, ($ INV) fails, and quantum gravity violates CPT invariance.[28]

The demonstration is simple. If $ takes pure states to mixtures, $\$^{-1}$ will take some non-trivial mixture W to a pure state $|\alpha><\alpha|$. $\$^{-1}$ should be linear; only states, which are positive operators, should lie in its range. From this it follows (see Wald [1980], p. 2749, for details) that $\$^{-1}$ can take a non-trivial mixture W to a pure state $|\alpha><\alpha|$ only if, for each pure state $|\varphi_i><\varphi_i|$ in W's spectral resolution,

$$\$^{-1}|\varphi_i><\varphi_i| = |\alpha><\alpha| \tag{A1}$$

Let $ act on each side of this equation to obtain

$$\$|\alpha><\alpha| = |\varphi_i><\varphi_i| \tag{A2}$$

But if there is more than one i (which there is, because we assumed W to be a non-trivial mixture), (A2) is impossible. Assuming $ to be invertible generates a contradiction. $ is not invertible, ($ INV) fails, and quantum gravity is not CPT invariant.

How then can Page offer $ which both oversees pure-to-mixed state evolution and is 'invertible within the restricted set of density matrices comprising its range' (Page [1994], p. 4)? Simply by meaning something idiosyncratic by 'invertible.' Consider his sample $. The statistical state of a spin ½ system is represented by a 2×2 density matrix ρ. But this statistical state can also be characterized by a polarization vector **P**, whose three components, plus the requirement that the diagonal elements of ρ sum to 1, determine ρ's components. Page concocts a superscattering matrix $ whose effect is to multiply the polarization vector of the in state by a real number $\lambda \in [0,1]$ to obtain the polarization vector of the out state. If $\lambda \neq 1$, only mixed states lie in the range of this $. So if $\lambda \in [0,1]$, non-unitary $ will issue a final mixed state which uniquely determines an initial state through the following prescription: take the polarization vector of the out state and divide by λ to obtain the polarization vector of the in state. Page's proposal is to use this inversion procedure to define an 'inverse' of $ *on the set of states whose preimage by $ lies in* $D(\mathcal{H}_{in})$. Because precisely these states are possible endpoints of evolution described by $, a $\$^{-1}$ so restricted suffices for retrodiction. So long as $\mathcal{H}_{\text{in}}$ and

[28] Wald's result is just the contraposition of Page ([1980]), which contends that any black hole evaporation process which can be described by a CPT-invariant superscattering matrix can be described by an S matrix mapping pure states to pure states. Page initially took the failure of CPT invariance to be a reason for rejecting Hawking's conclusion of a pure-to-mixed transition. The S matrix Page discusses works in consort with a Hilbert space model of the 'hidden states' available a black hole interior 'consisting of a single state' ([1980], p. 302). This model requires the black hole to form in the same state—the lone state its Hilbert space allots it—regardless of the state of the matter collapsing to form it. So Page's purity-preserving S matrix is perhaps the earliest (and most sincerely offered) example of quantum bleaching!

$\mathcal{H}_{out}$ are of equal dimension, a nonunitary \$ with a restricted 'inverse' permitting failsafe retrodiction will be available (Page [1994], p. 4).

Restricted, $\$^{-1}$ is not a genuine inverse. $\$\$^{-1}$ is not everywhere defined, and so not the identity. To strip the shudder quotes from '\$'s ''inverse''', Page must extend its action to all elements of $D(\mathcal{H}_{out})$. But extending $\$^{-1}$ to pure states which do not lie in the range of \$ ends unhappily. Applying Page's inversion prescription to such pure states yields matrices with negative eigenvalues: outside its intended domain, $\$^{-1}$ fails to be a map from states to states. So Page's \$ is not a counterexample to Wald's argument, which concerns an \$ everywhere defined as a map from states to states. What's more, it's easy to see that Page's \$ cannot arise from a CPT invariant fundamental theory. For \$ fails (\$ INV), which requires (W, W′) to be a possible universe for \$ iff (θ^{-1} W′, θW) is. If (W, W′) is a possible universe for \$, with W a pure state—and this Page allows—and if θ is a (CP)T inversion operator, θW is also pure. But Page countenances only mixed states as endpoints of \$ evolution. So θW is not a possible final state of the universe, (θ^{-1} W′, θW) is not a possible universe for \$, and (\$ INV) fails. Predictability would be lost to observers adrift in a sea of Hawking radiation, unable to determine the state of their universe before black hole formation. Page purchases these observers retrodictive success in the face of a noninvertible \$ at the cost of positing a radical asymmetry of physically possible initial and final conditions, an asymmetry which violates CPT invariance.[29]

Appendix 2: Quantum field theory for non-globally hyperbolic spacetimes

The argument in Section **3** for pure-to-mixed state transitions in black hole evaporation assumed that a sensible QFT can be applied to non-globally hyperbolic spacetimes such as that of Figure 2. In this section we indicate why this assumption is probably safe, at least for the simple case of a free field obeying the Klein–Gordon equation. Yurtsever ([1994]) has given a construction which applies to non-globally hyperbolic spacetimes M, g_{ab} and which, given a space of global solutions to the Klein–Gordon equation on M, produces an algebra of fields with a subalgebra A(U) associated with an open region U ⊂ M. In general, this field algebra will not have very attractive properties. However, Yurtsever shows that if M, g_{ab} is 'micro-causal'[30] with respect to the Klein–Gordon equation and if U, V ⊂ M are relatively spacelike open sets, then the subalgebras A(U) and A(V) commute. Micro-causality can fail for spacetimes with closed timelike curves, but it can be met for spacetimes

[29] For more on Wald's view of fundamental symmetries in quantum gravity, see C. Liu ([1993]).
[30] See Yurtsever ([1994]) for details. These microcausality conditions guarantee that, on a local level, the propagation of the quantum field has nice properties.

such as that of Figure 2, where global hyperbolicity fails not because of closed timelike curves but because of naked singularities. Thus, our invocation in Section **3** of the commutation condition seems reasonable.

Kay ([1992]) has recommended a locality condition that goes beyond Yurtsever's micro-causality. A spacetime M, g_{ab} is said to be F-quantum compatible if it admits a global algebra satisfying the F-locality condition. The latter means (roughly) that each point of M should have a globally hyperbolic neighbourhood on which the standard algebraic structure of observables coincides with the structure induced by the global algebra. Kay, Radzikowski, and Wald ([1997]) have shown that acausal spacetimes whose chronology horizons are compactly generated are not F-local.[31] But again there seems to be no problem for spacetimes such as that of Figure 2 meeting F-locality.

The Yurtsever construction does not solve (nor does it pretend to solve) the difficult interpretational problems connected with naked singularities. In particular, the construction starts with a space of global solutions; but that choice depends on the boundary conditions that one expects or hopes the naked singularities to satisfy. Fortunately, we do not need to solve any of these problems; all we need is the assurance that whatever the choice of the space of global solutions and whatever the resulting algebra of observables, it has properties nice enough for our Lemma to be applied.

[31] Intuitively the chronology horizon separates the portion of spacetime in which there are closed timelike curves from the portion that contains them. The generators of this horizon are null geodesics. If, when traced backwards in time, these generators enter and remain in a compact set, the horizon is said to be compactly generated. Such a spacetime is a candidate for describing the operation of a time machine.

References

Albert, D. [1992]: *Quantum Mechanics and Experience*, Cambridge: Harvard University Press.

Banks, T. [1995]: 'Lectures on Black Holes and Information Loss', *Nuclear Physics Proceedings Supplement*, **41**, pp. 21–65.

Banks, T. and Susskind, L. [1984]: 'Difficulties for the Evolution of Pure States into Mixed States', *Nuclear Physics B*, **244**, pp. 125–34.

Bekenstein, J. B. [1993]: 'How Fast Does Information Leak Out from a Black Hole?' *Physical Review Letters*, **70**, pp. 3680–3.

Callan, C. G., Giddings, S. B., Harvey, J. A., and Strominger, A. [1992]: 'Evanescent Black Holes', *Physical Review D*, **45**, pp. R1005–9.

Clifton, R., Feldman, D. V., Redhead, M. L. G., and Wilce, A. [1997]: 'Superentangled States', quant-ph/9711020 (to appear in *Physical Review A*).

Coleman, S., Preskill, J., and Wilczek, F. [1992]: 'Quantum Hair on Black Holes', *Nuclear Physics B*, **378**, pp. 175–246.

Cornish, N. J. and Moffat, J. W. [1994]: 'Nonsingular Gravity without Black Holes', *Journal of Mathematical Physics*, **35**, pp. 6628–43.

Danielsson, U. H. and Schiffer, M. [1993]: 'Quantum Mechanics, Common Sense, and the Black Hole Information Paradox', *Physical Review D*, **48**, pp. 4779–84.

Earman, J. [1995]: *Bangs, Crunches, Whimpers and Shrieks: Singularities and Acausalities in Relativistic Spacetimes*, New York: Oxford University Press.

Ellis, J., Mavromatos, N. E., and Nanopoulos, D. V. [1991]: 'Quantum Coherence and Two-dimensional Black Holes', *Physics Letters B*, **267**, pp. 465–74.

Flam, F. [1993]: 'Plugging a Cosmic Information Leak', *Science*, **259**, pp. 1824–5.

Geroch, R. P. [1967]: 'Topology in General Relativity', *Journal of Mathematical Physics*, **8**, pp. 782–6.

Giddings, G. [1992]: 'Black Holes and Massive Remnants', *Physical Review D*, **46**, pp. 1347–52.

Giddings, G. [1993]: 'Black Holes and Quantum Predictability', hep-th/9306041.

Giddings, G. [1995]: 'Quantum Mechanics of Black Holes', in E. Gava, A. Masiero, K. S. Narain, S. Randjbar-Daemi, and Q. Shafi (*eds*), *1994 Summer School in High Energy Physics and Cosmology. ICTP Series in Theoretical Physics-Vol.11*, Singapore: World Scientific, pp. 530–74.

Hartle, J. B. [1995]: 'Spacetime Quantum Mechanics and the Quantum Mechanics of Spacetime', in B. Julia and J. Zinn-Justin (*eds*), *Gravitation and Quantizations: Les Houches Session LVII*, Amsterdam: Elsevier, pp. 285–480.

Hartle, J. B. [1998]: 'Generalized Quantum Theory in Evaporating Black Hole Spacetimes', in R. Wald (*ed.*), *Black Holes and Relativistic Stars*, Chicago: University of Chicago Press, pp. 195–219.

Hawking, S. W. [1975]: 'Particle Creation by Black Holes', *Communications in Mathematical Physics*, **43**, pp. 199–220.

Hawking, S. W. [1976]: 'The Breakdown of Predictability in Gravitational Collapse', *Physical Review D*, **14**, pp. 2460–73.

Hawking, S. W. [1982]: 'The Unpredictability of Quantum Gravity', *Communications in Mathematical Physics*, **87**, pp. 395–415.

Hawking, S. W. [1998a]: 'Is Information Lost in Black Holes?' in R. Wald (*ed.*), *Black Holes and Relativistic Stars*, Chicago: University of Chicago Press, pp. 221–40.

Hawking, S. W. [1998b]: 'Loss of Information in Black Holes', in S. Huggett, L. Mason, K. Tod, S. Tsou, and N. Woodhouse (*eds*), *The Geometric Universe: Science, Geometry, and the Work of Roger Penrose*, Oxford: Oxford University Press, pp. 123–33.

Hawking, S. W. and Ellis, G. F. R. [1973]: *The Large Scale Structure of Spacetime*, Cambridge: Cambridge University Press.

Hawking, S. W. and Stewart, J. M. [1993]: 'Naked and Thunderbolt Singularities in Black Hole Evaporation', *Nuclear Physics B*, **400**, pp. 393–415.

Helfer, A. D. [1996]: 'The stress-energy operator,' *Classical and Quantum Gravity*, **13**, L129–L134.

Horowitz, G. T. [1997]: 'Quantum States of Black Holes', gr-qc/9704072.

Johnson, J. [1998]: 'Physical Laws Collide in Black Hole Event', *New York Times*, 7 April 1998.

Kay, B. [1992]: 'The Principle of Locality and Quantum Field Theory on (Non Globally

Hyperbolic) Spacetimes', *Reviews of Mathematics Physics, Special Issue*, pp. 167–95.

Kay, B., Radzikowski, M. J., and Wald, R. M. [1997]: 'Quantum Field Theory on Spacetimes with a Compactly Generated Cauchy Horizon', *Communications in Mathematical Physics*, **183**, pp. 533–56.

Kuchař, K.V., Romano, J. D., and Varadarajan, M. [1997]: 'Dirac Constraint Quantization of a Dilatonic Model of Gravitational Collapse', *Physical Review D*, **55**, pp. 795–808.

Liu, C. [1993]: 'The Arrow of Time in Quantum Gravity', *Philosophy of Science*, **60**, pp. 619–37.

Lowe, D. A., Susskind, L., and Uglum, J. [1994]: 'Information Spreading in Interacting String Theory', *Physics Letters B*, **327**, pp. 226–33.

Moffat, J. W. [1993]: 'Do Black Holes Exist?' gr-qc/9302032.

Myers, R. [1997]: 'Pure States Don't Wear Black', *General Relativity and Gravitation*, **29**, pp. 1217–22.

Page, D. [1980]: 'Is Black-Hole Evaporation Predictable?' *Physical Review Letters*, **44**, pp. 301–4.

Page, D. [1994]: 'Black Hole Information', in R. B. Mann and R. G. McLenaghan (*eds*), *Proceedings of the 5th Canadian Conference on General Relativity and Relativistic Astrophysics*, Singapore: World Scientific, pp. 1–41.

Parikh, M. and Wilczek, F. [1998]: 'Global Structure of Evaporating Black Holes', gr-qc/9807031.

Polchinski, J. and Strominger, A. [1994]: 'Possible Resolution of the Black Hole Information Puzzle', *Physical Review D*, **50**, pp. 7403–9.

Preskill, J. [1991]: 'Quantum Hair', *Physica Scripta*, **T56**, pp. 258–64.

Preskill, J. [1993]: 'Do Black Holes Destroy Information?', in S. Kalara and D. V. Nanopoulos (*eds*), *Black Holes, Membranes, Wormholes and Superstrings*, Singapore: World Scientific, pp. 23–39.

Russo, J., Susskind, L., and Thorlacius, L. [1993]: 'Endpoint of Hawking Evaporation', *Physical Review D*, **46**, pp. 3444–9.

Schiffer, M. [1993]: 'Is it possible to recover information from the black-hole radiation?' *Physical Review D*, **48**, pp. 1652–8.

Srednicki, M. [1993]: 'Is Purity Eternal?' *Nuclear Physics B*, **410**, pp. 143–54.

Strominger, A. [1996]: 'Lectures on Black Holes', in F. David, P. Ginsparg, and J. Zinn-Justin (*eds*), *Fluctuating Geometries in Statistical Mechanics and Field Theory: Les Houches Session LXII*, New York: Elsevier, pp. 699–761.

Susskind, L. [1994]: 'Strings, Black Holes, and Lorentz Contraction', *Physical Review D*, **49**, pp. 6606–11.

Susskind, L. [1997]: 'Black Holes and the Information Paradox', *Scientific American*, **272**, 4, April, pp. 52–7.

Susskind, L. and Thorlacius, L. [1994]: 'Gedanken Experiments Involving Black Holes', *Physical Review D*, **49**, pp. 966–74.

Susskind, L., Thorlacius, L., and Uglum, J. [1993]: 'The Stretched Horizon and Black Hole Complementarity', *Physical Review D*, **48**, pp. 3743–61.

Takesaki, M. [1979]: *Theory of Operator Algebras I*, New York: Springer Verlag.

't Hooft, G. [1985]: 'On the Quantum Structure of a Black Hole', *Nuclear Physics B*, **256**, pp. 727–45.

't Hooft, G. [1990]: 'The Black Hole Interpretation of String Theory', *Nuclear Physics B*, **335**, pp. 138–54.

't Hooft, G. [1997]: 'Distinguishing Causal Time from Minkowski Time and a Model for the Black Hole Quantum Eigenstates', gr-qc/9711053.

Torre, C. and Varadarajan, M. [1998]: 'Functional evolution of free quantum fields,' hepth/981122.

Unruh, W. G. and Wald, R. M. [1995]: 'Evolution Laws Taking Pure States to Mixed States in Quantum Field Theory', *Physical Review D*, **52**, pp. 2176–82.

Visser, M. [1998]: 'Hawking Radiation without Black Hole Entropy', *Physical Review Letters*, **80**, pp. 3436–9.

Wald, R. M. [1980a]: 'Quantum Gravity and Time Reversibility', *Physical Review D*, **21**, pp. 2742–55.

Wald, R. M. [1980b]: 'Dynamics in Nonglobally Hyperbolic, Static Spacetimes', *Journal of Mathematical Physics*, **21**, pp. 2802–5.

Wald, R. M. [1984a]: *General Relativity*, Chicago: University of Chicago Press.

Wald, R. M. [1984b]: 'Black Holes, Singularities and Predictability', in S. M. Christenson (*ed.*), *Quantum Theory of Gravity*, Bristol: Adam Hilger, pp. 160–8.

Wald, R. M. [1994]: *Quantum Field Theory in Curved Spacetimes and Black Hole Thermodynamics*, Chicago: University of Chicago Press.

Yurtsever, U. [1994]: 'Algebraic Approach to Quantum Field Theory on Non-globally Hyperbolic Spacetimes', *Classical and Quantum Gravity*, **11**, pp. 999–1012.