PHILOSOPHICAL TRANSACTIONS A

rsta.royalsocietypublishing.org

Research

Article submitted to journal

Quantum information vs. black hole physics: Deep firewalls from narrow assumptions

Samuel L. Braunstein & Stefano Pirandola

Computer Science, University of York, York YO10 5GH, United Kingdom

The prevalent view that evaporating black holes should simply be smaller black holes has been challenged by the firewall paradox. In particular, this paradox suggests that something different occurs once a black hole has evaporated to one-half its original surface area. Here we derive variations of the firewall paradox by tracking the thermodynamic entropy within a black hole across its entire lifetime and extend it even to AdS spacetimes. Our approach sweeps away many unnecessary assumptions, allowing us to demonstrate a paradox exists even *after* its initial onset (when conventional assumptions render earlier analyses invalid). The most natural resolution may be to accept firewalls as a real phenomenon. Further, the vast entropy accumulated implies a *deep firewall* that goes "all the way down" in contrast to earlier work describing only a structure at the horizon.

The fundamental physics of black holes has been an enduring mystery [1]. Great progress has been recently made with the discovery of the firewall paradox for black holes, which suggests the existence of a manifestly strong phenomenon, the *firewall* [2] or *energetic curtain* [3] as it was originally dubbed in 2009. It is generally assumed that there exists a correspondence between the physical characteristics of a real (i.e., quantum mechanical) black hole and its theoretical classical counterpart. Loosely, the firewall paradox constructs a contradiction between this correspondence, the black hole's thermodynamic behavior, and quantum unitarity. This presents a fundamental opportunity to reconcile gravity and quantum mechanics.

THE ROYAL SOCIETY PUBLISHING

Here, we derive a firewall paradox with many unnecessary assumptions removed: We do not rely on an ability to measure or decode the Hawking radiation. Indeed, the computational complexity of this latter task has been claimed [4] to render earlier arguments supporting the firewall paradox [2] as fatally flawed. Nor do we rely on any specific radiation process like pair creation or tunneling. Indeed, if firewalls were real, then 'nice time slices' through a black hole's spacetime [5] and all mechanisms associated with them could no longer be trusted. Further, black hole evaporation is strongly *believed* to be non-local (e.g., via the holographic principle [6,7]). However, pair creation in particular comes from *local* quantum field theory [1]. Earlier derivations of the paradox [2] are therefore apparently fraught with difficulties.

We go on to explore the role of non-local physics across the horizon with a second theorem. Our analysis points to an extra assumption left out of conventional holographic approaches [6–8]. Our two paradoxes provide insight into the physics of black holes across their entire lifetime. Evidence supporting our assumptions and variations to them are given in the discussion.

We will show that the unitary evaporation of a black hole leads to an excess of entropy in the vicinity of the black hole, peaking when the black hole surface area has shrunk by a factor of two. This excess is huge, e.g., for a single stellar mass black hole it is the roughly the entropy contained in all the stars in the observable universe [9]. Even a small fraction of this excess entropy existing external to the event horizon would fundamentally change the observable characteristics of a black hole. We argue in the discussion that the only *feasible* alternatives: via Theorem 1, are the existence of a firewall with this excess entropy literally filling the black hole interior; and independently via Theorem 2, that no black hole 'horizon' can act as a 'surface of no return'.

Thermodynamic entropy from remote entanglement: Our thermodynamic analysis will rely on the von Neumann entropy $S(X)$ of a system X and the quantum mutual information $I(X:Y) \equiv S(X) + S(Y) - S(X,Y)$ which quantifies the correlations between a pair of systems X and Y.

Suppose X and Y are *remotely* separated and non-interacting systems whose correlations correspond to maximally entangled subspaces within X and Y. Then this subspace of X must have Schmidt coefficients consistent with the entanglement and hence carrying entropy $\frac{1}{2}I(X{:}Y)$. By mimicing this subspace with a locally created probabilistic state given by $\mathrm{tr}_Y(\rho_{XY})$ every local property, including thermodynamic, can be seen to be identical in the original and mimiced states. Thus, there will be an amount of thermodynamic entropy in X (and in Y) at least equal to $\frac{1}{2}I(X{:}Y)$ which will be observable by even highly delocalized detectors. As an aside, we note that this is in contrast to local correlations due, for example, to entanglement in the vacuum state across a boundary [10], which would be unobservable (and certainly not thermodynamic in nature) unless one's (inertial) detectors were localized on scales comparable to the cutoff (presumably Planckian).

Entropic bounds for the black hole neighborhood: The standard correspondence between real and classical black holes [1] assumes that black holes evaporate into vacuum (as seen by an infalling observer). Our two theorems shall not assume any particular quantum state (vacuum, near-vacuum or otherwise) for the quantum field into which the black hole evaporates. We shall replace this with the much weaker assumption of a *non-exotic atmosphere*. In part, we achieve this by focusing on the gross thermodynamic properties of the neighborhood, N, external to and surrounding a black hole that reaches out far enough to encompass any process by which radiation is produced.

Let us recall 't Hooft's entropic bound on ordinary matter [6]: It shows that if one excludes configurations of ordinary matter that will inevitably undergo gravitational collapse, one finds $\mathcal{A}^{3/4} \geq S_{\text{matter}}$, where $\mathcal{A}$ is the surface area of the region containing the matter (in Planck units) and S_{matter} is the thermodynamic entropy of the matter (with k_B set to unity).

Suppose that the external neighborhood of a black hole, at some specific epoch, has a surface area μ times that of the black hole's horizon, so $\mathcal{A}_N = 4\mu\, S_{\text{BH}}$, where S_{BH} denotes the black hole's Bekenstein-Hawking entropy. Then to satisfy 't Hooft's entropic bound on ordinary matter [6,11],

the thermodynamic entropy of the external neighborhood, $S^N_{\rm therm}$, must be bounded above by

$$\mathcal{A}_N^{3/4} \equiv \left[2\sqrt{2}\left(\frac{\mu^3}{S_{\rm BH}}\right)^{1/4}\right] S_{\rm BH} \geq S^N_{\rm therm}. \tag{0.1}$$

For large black holes, the prefactor in square brackets is much much less than unity even for extremely large neighborhoods, e.g., $\mu = 10^4$. If this bound fails, the external neighborhood, N, *must* consist of some exotic matter, e.g., an atmosphere of microscopic black holes.

Theorem 1: A contradiction exists among: 1.a) complete and unitary black hole evaporation, 1.b) freely falling observers notice nothing special until well within a large black hole's horizon, and 1.c) the black hole interior Hilbert space dimensionality may be well approximated as the exponential of the Bekenstein-Hawking entropy.

Proof: Assumption 1.a has two ingredients (i and ii): that black hole evaporation is unitary and that it is complete, respectively, leaving behind no remnant. We start by unraveling the implications of these two ingredients separately:

Unitary evaporation 1.a(i): Consider the unitary generation of radiation from a black hole by an arbitrary process, Fig. 1. We associate this process with some specific black hole and presume that to an excellent approximation radiation is not produced beyond N.

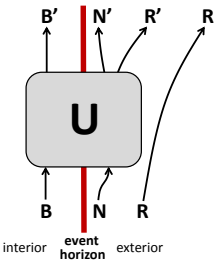

Figure 1. Schematic generation of radiation R' during some epoch by an arbitrary unitary process U. Here, N and N' are the degrees-of-freedom in the black hole's exterior neighborhood prior and posterior to the unitary operation, respectively; B and B' label interior degrees-of-freedom; and R denotes radiation from an earlier epoch. Note, the initial joint quantum state of (B, N, R) is arbitrary.

As with the original firewall paradox [2] we rely on strong subadditivity of the quantum mutual information: $I(W, X{:}Y) \geq I(X{:}Y)$ for any extra system W. Applying strong subadditivity to Fig. 1 yields $I(B', N', R'{:}R) \geq I(R'{:}R)$. Moreover, as entropy is invariant under unitary operations $I(B', N', R'{:}R) = I(B, N{:}R)$, we find $I(B, N{:}R) \geq I(R'{:}R)$, which eliminates B' and N', allowing us to work instead with B and N that are associated with the earlier-time black hole.

For simplicity, consider a black hole created by collapsing matter initially in a pure quantum state (the result holds in general, see appendices). In this case, the global state of (B, N, R) may also be treated as pure implying $S(B, N, R) = 0$, $S(B, N) = S(R)$, $S(B, R) = S(N)$, $S(N, R) = S(B)$ and hence $I(B, N{:}R) = I(B{:}R) + I(N{:}R)$ which is therefore bounded below by

$$I(B : R) + I(N : R) \geq I(R' : R). \tag{0.2}$$

Complete evaporation 1.a(ii): After the pure state black hole has completely evaporated away the net radiation, (R, R'), should also be in a pure quantum state to preserve unitarity, so that $S(R', R) = 0$. Thus one might expect that correlations would exist between the early and late epoch radiation, R and R', respectively.

In fact, the study of random unitary operations allows us to say much more: Since the Hilbert space dimensionalities involved are so huge Levy's lemma guarantees a generic behavior for entropy in all but a set of measure zero of evaporation mechanisms [3]. (Indeed Levy's lemma can be interpreted as implying that mechanisms with behavior different from the generic will be exponentially unstable to any but a set of measure zero of perturbations.)

The consequences of random unitary evaporation [3,12] may be succinctly summarized: The entropy of the radiation grows monotonically (at almost exactly the maximal rate of one bit's

worth of entropy per qubit of radiation emitted) until the *Page time*, when the black hole's area has halved. From the Page time (PT) onward the overall entropy in the radiation decreases at the same rate, reaching zero when evaporation is complete [3,12]. (Note, that the terms 'bits' and 'qubits' are used as units for information content and do not imply two-level systems.) Thus, combining both the pre- and post-PT results we may write $S(R) = S(R') = \min(\ln|R|, \ln|R'|)$, where $|X| = \dim(X)$.

The Bekenstein-Hawking entropy quantifies the thermal entropy in a black hole from the first law of black hole thermodynamics. This presumes, however, that the radiation lacks correlations. Thus by ignoring the correlations above we may identify [13] the Bekenstein-Hawking entropy of the *initial* black hole as $S_{\rm BH} = \ln|R| + \ln|R'|$. Similarly, after evaporating radiation R, the remaining black hole's Bekenstein-Hawking entropy is $\ln|R'|$. Labeling $|R| < |R'|$ and $|R| > |R'|$ as pre- and post-PT, respectively, then gives

$$\tfrac{1}{2}\, I(R' : R) = \min(\ln|R|, \ln|R'|) = \begin{cases} S_{\rm BH} - \ln|R'|, & \text{pre-PT}, \\ \ln|R'|, & \text{post-PT}. \end{cases} \tag{0.3}$$

As only extreme violations of assumption 1.b are of relevance here, it is convenient to decompose it into (i) external and (ii) internal constraints

External free-fall equanimity 1.b(i): An exceedingly weak assumption is that a freely-falling observer notices no exotic matter at least down to the horizon. Thus, from Eq. (0.1), the radiation-correlated contribution to the neighborhood's entropy $\frac{1}{2} I(N{:}R) \leq S^{N}_{\rm therm}$ must be negligible. Combining this with Eq. (0.2) yields

$$\tfrac{1}{2}\, I(B : R) \gtrsim \tfrac{1}{2}\, I(R' : R). \tag{0.4}$$

Finite interior Hilbert space 1.c: In order to ensure that the black hole interior (within the stretched horizon) can contain no more physical entropy than can eventually evaporate away, it is assumed that the interior has a Hilbert space dimensionality that is well approximated by the exponential of the Bekenstein-Hawking entropy. This implies that $\ln|B| \simeq \ln|R'|$ during a black hole's evaporation and hence from Eqs. (0.3) and (0.4) we have

$$\ln|B| \geq \tfrac{1}{2} I(B : R) \gtrsim \begin{cases} S_{\rm BH} - \ln|B|, & \text{pre-PT}, \\ \ln|B|, & \text{post-PT}. \end{cases} \tag{0.5}$$

Initially, the black hole interior accumulates thermodynamic entropy at a rate of (at least) one bit per qubit radiated. From the PT onward, the shrinking interior (everything within the stretched horizon) is filled with half of a maximally entangled state, $\frac{1}{2} I(B{:}R) \simeq \ln|B|$, corresponding formally to an infinite-temperature thermal state.

Internal free-fall equanimity 1.b(ii): A contradiction ensues if we assume that an infalling observer notices nothing, or at worst a low entropy state, well within the horizon. Indeed, past the PT an observer inside the black hole (within the stretched horizon) can no longer escape intimate contact with an infinite temperature. ■

This result follows straightforwardly and differs from earlier arguments [2] that invoke a likely impossible capacity for decoding Hawking radiation [4]. In the appendices we extend our analysis to consider mixed state black holes; include the thermodynamic entropy in N; and exclude contributions of Planck-scale black holes.

The AdS/CFT correspondence: The strongest contender for a fully unitary theory of black hole evaporation involves the AdS/CFT correspondence [8], which formalizes the holographic principle [6,7]. Notably, this theory fails to predict a firewall (or anything unusual within the horizon) from the PT onward.

Provided AdS black holes evaporate, Theorem 1 straightforwardly extends to this scenario and to AdS black holes more generally (see appendices): Initially, there is *no* firewall; however, by the PT the paradox is reinstated.

The AdS/CFT correspondence champions non-local physics, but remains silent as to why the (local) 'nice time slice' argument is apparently wrong (even prior to the PT). Below we rigorously

reevaluate the role of non-locality during black hole evaporation. We prove a contradiction would arise between unitarity and locality, only when a third assumption holds. Our analysis makes no recourse to holographic reasoning [6–8].

Theorem 2: A contradiction exists among: 2.a) complete and unitary black hole evaporation, 2.b) large black holes are described by local physics and 2.c) externally, a large black hole resembles its classical counterpart (aside from its slow evaporation).

The proof is similar to that of Theorem 1 (see appendices).

Discussion: Both theorems apply to the behavior of large black holes where General Relativistic reasoning is conventionally expected to hold. Theorem 1 yields a contradiction from the Page time onward, suggesting that huge thermodynamic entropies reside within the black hole. Theorem 2 incorporates the local structure of a black hole's horizon and yields a contradiction almost immediately once evaporation has begun, suggesting a failure of strict locality across the horizon.

Let us first consider rejecting assumption 1.a (2.a). Unfortunately, any loss of unitarity associated with the physics of black holes would likely infect every quantum mechanical process [6]. Similarly, if black hole evaporation were not complete, the infinite degeneracy of the resulting 'stable remnant' would cause all virtual loops of such remnants to produce infinite cross sections, again causing a general loss of unitarity [6,14,15]. Thus, rejecting assumption 1.a (2.a) appears to be unacceptable.

If we accept 1.a (2.a), either theorem leaves us with a striking dichotomy. Let us start with the consequences of Theorem 1. We must reject at least one assumption of 1.c or 1.b. To start with, were the accessible dimensionality within the stretched horizon larger than the estimate given by the Bekenstein-Hawking entropy (e.g., see Ref. [16]), we would be able fill a black hole with more thermodynamic entropy than could be accounted for by the entropy that would eventually appear as radiation. Thus, a failure of assumption 1.c is tantamount to a death sentence for the theory of black hole thermodynamics, and possibly thermodynamics itself.

By contrast, assumption 1.b, although usually considered a consequence of the Equivalence Principle of General Relativity is no more than a boundary condition on the quantum fields at the horizon; it is well known that different choices of 'vacuum state' lead to wildly different behaviors for the energy-momentum tensor there. Splitting 1.b into its components: A failure of 1.b(i) would imply that the exterior must consist of super-entropic exotic matter (such as an atmosphere of microscopic black holes), and so would almost certainly have some observational consequences. Finally, a failure of 1.b(ii) would imply that by the Page time a black hole's interior is *filled* with half of a maximally entangled state.

In 2009 it was noted that maximal entanglement between the black hole interior and the radiation implied an absence of entanglement across the horizon [3]: "In an arbitrary system where trans-boundary entanglement has vanished, the quantum field cannot be in or anywhere near its ground state. Applied to black holes, a loss of trans-event horizon entanglement implies fields far from the vacuum state in the vicinity of the event horizon." We now extend this reasoning: As there is *no* entanglement between *any* pieces of the quantum fields within the black hole, no place within the interior can look like a low-energy state — like regular spacetime.

Next, consider the options left by Theorem 2: to reject at least one of the assumptions 2.c or 2.b. Rejecting 2.c is equivalent to rejecting 1.b(i). Therefore, the only other minimal option (rejecting assumption 2.b) would be to assume that communication from the black hole interior to exterior across the horizon was possible. In particular, one might note that a "tunneling" mechanism has been long anticipated to provide a more powerful explanation for black hole radiation [3,13,17]. However, tunneling across the horizon alone [17] is insufficient, as it still leaves unanswered how the degrees-of-freedom from deep within a black hole manage to (non-locally) reach up to just inside the horizon where they can participate in such tunneling.

Simplifying our assumptions: It turns out that there are several ways of reducing the assumptions needed to obtain *Theorem 1*. First, we may drop assumption 1.c entirely if we can apply the holographic entropy bound [18] within a black hole's horizon. Recall the converse interpretation of this bound which states that the minimal area encompassing a given thermodynamic entropy

is four times that entropy (in Planck units). The covariant form of this bound [18] should apply anywhere (including within a black hole's horizon). Hence, let $\mathcal{A}_{\rm min}$ be the minimal area encompassing a thermodynamic entropy of $\frac{1}{2}I(B{:}R)$. By assumption 1.b, this area must be well within the (surface area of the) horizon, implying that $\ln|R'| \gg \frac{1}{4}\mathcal{A}_{\rm min} \geq \frac{1}{2}I(B{:}R)$, which directly contradicts Eqs. (0.3) and (0.4). This bound shows that from the Page time, an observer freely falling from infinity will hit the firewall almost immediately upon crossing the horizon.

Ineluctable modality of the invisible: Holographic entropy bounds [18] show that space is a container of sorts. A region of space cannot contain more thermodynamic entropy than one-quarter of its bounding area. That container can be empty, full, or anywhere in-between.

The conventional view of a black hole is that its interior is effectively empty of entropy (devoid of matter of any kind). The firewall paradigm, as we have unraveled here, is that the interior will be slowly filled up with entropy by any generic evaporation process.

Unlike previous work [2,3], we have shown that the firewall is not merely signaled by a loss of entanglement across the event horizon, but across all surfaces internal to the black hole. However, entanglement is necessary for any kind of low-energy behavior of quantum fields on a manifold to be possible [3]. In this sense, one might say that spacetime ceases to exit within a developed firewall.

Such a "deep firewall" structure is naively at odds with the internal description of text-book vacuum solutions for black holes. Structures analogous to our deep firewall have been recently found in self-consistent solutions to the semiclassical Einstein field equations for black holes at thermal equilibrium with a heat bath [19]. One might question the legitimacy of such calculations due to their limitation to Schwarzschild-like black holes, their requiring a violation of the dominant energy condition, their unphysical environment, or the detailed assumptions used to constrain the energy-momentum tensor. Nevertheless, our work here implies that deep firewalls are a universal feature of old black holes of any type undergoing evaporation by any unitary process.

It remains to speculate about whether the event horizon survives the formation of a deep firewall just behind it or is instead replaced by it. In the former case, evaporation would continue by something very much like quantum tunneling [3,13,17] from degrees-of-freedom just inside the horizon. In the latter case, evaporation may continue via direct ejection from the *entanglar* (entangled star) though its detailed spectrum (e.g., its neutrino flux) and lifetime would almost certainly differ from a true black hole with otherwise identical mass, charge and angular momentum. Naively, an entanglar (of even a modest size) would take far longer than the age of the universe to evolve from a black hole, so none can be expected to currently exist. Conversely, the unambiguous observation of such an entanglar would yield *prima facie* evidence for an object that far predates the Big Bang.

APPENDICES

Theorem 1 for Anti de Sitter space

Clearly theorem 1 is stated in very general terms, however, one might worry that there are implicit assumptions about the global (asymptotic) geometry within which the black hole of interest sits. In fact, all that is really required is that the radiation can be remotely separated from the black hole's interior and its surrounding neighborhood (on distances large compared to the UV cutoff, presumably the Planck scale). This is all we need to ensure that maximal entanglement between R and B and/or N may be identified as contributing to the thermodynamic entropy of these systems (as discussed in the main text).

What about if there is a boundary to the spacetime, such as in Anti de Sitter (AdS) space? Here the radiation could travel all the way to the boundary and return in a finite time, so we have to worry about any returning radiation which is no longer well separated. In fact, there is a simple accounting (or relabeling) trick that allows us to handle this straightforwardly. In order to see what this relabeling involves let us first recall that Levy's lemma constrains the correlations in a

very general way: In particular, the order in which subsystems are 'sampled' is irrelevant, only the dimensionalities of the identified subsystems plays a role. This means that we can simply redefine R (see Fig. 1) as that radiation from an earlier epoch that remains well separated from the black hole and its surrounding neighborhood at the epoch of interest. Any earlier radiation that falls back into the black hole neighborhood (or into the black hole itself) is then incorporated into B', N', R' or some combination of all three.

With that relabeling everything goes through as before. The 'Page time' (now no longer necessarily a time) refers to any *scenario* where the thermodynamic entropy of any radiation, from earlier epochs and which remains well separated from (B, N), equals one-half the Bekenstein Hawking entropy of the original black hole. The only 'obstacle' for obtaining the theorem then is for a scenario where the black hole is so large within the AdS space that there is never a situation where the Page time is reached because the radiation is returning too quickly from the AdS boundary.

Proof of Theorem 2

Proof: Assumptions 1.a and 2.a are identical.

Local physics 2.b: We shall suppose that however the radiation is generated, it is constrained to be local (at least for large black holes). Local physics forbids communication across light cones [20], and hence also from within a black hole's event horizon to the exterior. We now recall the no-communication decomposition theorem [21] (see Fig. 2) stating that any unitary process U which maps subsystems A and B into A' and B' but which does not allow communication from A to B', $A \not\rightsquigarrow B'$, can be decomposed into a pair of unitary subprocesses V and W connected by a 'reverse communication' channel C.

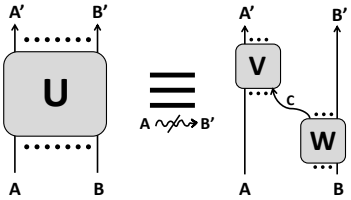

Figure 2. Quantum circuit for the no-communication decomposition theorem. Assuming $A \not\rightsquigarrow B'$, the unitary process on the left maps onto the pair of processes on the right.

Theorem 2 requires that the inputs and outputs form distinct components of a tensor product decomposition of the overall Hilbert space, a requirement which is automatic for finite dimensions. For any local quantum field theory we may rely on the commutativity of operators with support only outside each others light cones. Thus, locality dictates the existence of the required tensor product structure across a black hole's horizon [1,3]. Applying the circuit equivalence in Fig. 2 to the Fig. 1 gives the structure of an arbitrary unitary black hole evaporation process consistent with local physics (see Fig. 3).

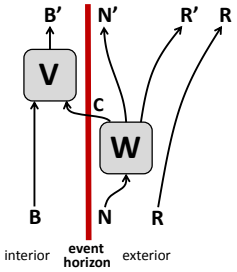

Figure 3. Schematic unitary generation of radiation from a black hole whose horizon obeys local physics, forbidding communication from interior to exterior (assumptions 2.a and 2.b, see text). Here C denotes possible 'reverse communication'.

Note that Fig. 3 is not to be interpreted as a spacetime diagram. In particular, we do not require that there is any space-like hypersurface which simultaneously cuts through the subsystems there displayed. For example, we do not require that subsystem C all arrives in one block for unitary processing inside the black hole. From this perspective, a quantum circuit is a powerful construct.

Strong subadditivity in Fig. 3 gives $I(C, N', R':R) \geq I(R':R)$, and using the unitary invariance of entropy we have $I(C, N', R':R) = I(N:R)$ leading to

$$I(N:R) \geq I(R':R). \tag{0.6}$$

Note that this inequality involves only correlations between external degrees-of-freedom and hence relates quantities which are, in principle, directly observable and reportable. Combining this with the assumption of complete evaporation, Eq. (0.3), we find

$$\tfrac{1}{2} I(N:R) \geq \begin{cases} S_{\rm BH} - \ln|R'|, & \text{pre-PT}, \\ \ln|R'|, & \text{post-PT}. \end{cases} \tag{0.7}$$

Non-exotic atmosphere 2.c: We shall take assumption 2.c to be equivalent to 1.b(i), that the exterior should not consist of exotic matter. The so-called holographic entropy bound [18] shows that an entropy of at most $O(\mathcal{A}_{\rm horizon}^{1/2})$ can reside between the causal and stretched horizons, so their distinction has negligible effect here. To ensure a non-exotic atmosphere for a partially evaporated black hole with Bekenstein Hawking entropy $\ln|R'|$, Eq. (0.1), implies $\ln|R'| \gg S_{\rm therm}^{N} \geq \frac{1}{2} I(N:R)$. Combining this with Eq. (0.7) gives

$$\ln|R'| \gg \tfrac{1}{2} I(N:R) \geq \begin{cases} S_{\rm BH} - \ln|R'|, & \text{pre-PT}, \\ \ln|R'|, & \text{post-PT}. \end{cases} \tag{0.8}$$

Except for the very earliest stages of evaporation (and well before the PT), this result yields a contradiction. ■

Firewall paradoxes including negligible entropies

In this section we repeat the key elements of both theorems with the following modifications: (a) We explicitly include the entropy in the atmosphere, bounding its size rather than merely considering it to be negligible; (b) We only follow the black hole evaporation to the point where the black hole is still much larger than Planck scale. To illustrate that neither of these changes affect the results of the main text we focus solely on the behavior at the Page time.

Theorem 1:

Consider now the scenario where we follow a black hole to a relatively late stage of its complete evaporation. In particular, when its area has shrunk to some small fraction of its original size, but is still much larger than the Planck scale so that the physics of Planck scale black holes plays no part. We denote all pre-Page time radiation as R and the post-Page time radiation as R' (produced up until the black hole has reached a specified fraction, say roughly $\varepsilon/2$, of its original area). It follows therefore from the generic behavior of entropy during evaporation [3] that

$$I(R':R) = (1-\varepsilon) S_{\rm BH}, \qquad \varepsilon \ll 1. \tag{0.9}$$

Combining this with Eq. (0.2) we find

$$I(B:R) + I(N:R) = I(B,N:R) \geq (1-\varepsilon) S_{\rm BH}, \ \varepsilon \ll 1. \tag{0.10}$$

Equation (0.10) tells us that the radiation is almost perfectly maximally entangled with a subspace of the joint system (B, N) and as R quickly becomes remotely separated we may conclude that $\frac{1}{2}(1-\varepsilon) S_{\rm BH}$ represents a lower bound to the thermodynamic entropy of this joint system.

Free-fall equanimity: Consider now a freely-falling observer who is believed to see nothing special until they pass well within a large black hole's horizon (assumption 1.b).

Externally, 1.b(i), we assume that our infalling observer is not passing through an atmosphere of exotic matter prior to reaching the horizon. Therefore from Eq. (0.1), we have $\frac{1}{2}I(N{:}R) \ll \frac{1}{2}S_{\rm BH}$ for a black hole at the Page time. Thus, there exists a parameter $\eta \ll 1$, such that $I(N{:}R) \leq \eta\, S_{\rm BH}$.

Internally, this implies that Eq. (0.10) reduces to

$$I(B:R) \geq (1-\varepsilon-\eta)S_{\rm BH}, \qquad \varepsilon,\eta \ll 1. \tag{0.11}$$

Now, a trivial bound to the quantum mutual information is that $2\ln(\dim(B)) \geq I(B{:}R)$. If this bound were saturated, the huge thermodynamic entropy inside B would imply that an infalling observer would *immediately* encounter an incredibly mixed state (e.g., a near uniform mixture of roughly $10^{10^{77}}$ orthogonal quantum states for an initially stellar mass black hole) with correspondingly huge energies as soon as they passed the horizon. They would immediately encounter an 'energetic curtain' [3] or firewall [2] upon entering the black hole. To guarantee assumption 1.b(ii) holds, the above dimensional bound must be far from saturation, i.e., at the Page time

$$\ln(\dim(B)) \gg \frac{1}{2}(1-\varepsilon-\eta)S_{\rm BH}, \qquad \varepsilon,\eta \ll 1. \tag{0.12}$$

Finite interior Hilbert space: We may now derive a contradiction along the lines of the original firewall paradox. Assumption 1.c holds that the black hole interior has a Hilbert space dimensionality that is well approximated by the exponential of the Bekenstein-Hawking entropy. Thus, at the Page time, when a black hole's surface area has shrunk to one-half of its original value we would have

$$\ln(\dim(B)) \simeq \frac{1}{2}\, S_{\rm BH}, \tag{0.13}$$

which directly contradicts Eq. (0.12). ■

Theorem 2:

Note that Eq. (0.6) involves only correlations between external degrees-of-freedom and hence relates quantities which are, in principle, directly observable and reportable. Combining this with the assumption of complete evaporation, Eq. (0.9), we easily find

$$I(N:R) \geq (1-\varepsilon)S_{BH}, \qquad \varepsilon \ll 1. \tag{0.14}$$

Locality (assumption 2.b) has allowed us to eliminate B from Eq. (0.10), which in turn allows us to do without any specific bound to the size of the interior Hilbert space. More surprisingly, locality implies a very different picture: one where huge thermodynamic entropies must reside outside the black hole instead of inside it.

At first sight, this appears reminiscent of arguments based on time-reversing Hawking radiation. Ordinary Hawking radiation evolves out of vacuum modes, but any (information bearing) deviations were argued to have started out as high-energy excitations near the horizon [22]. By contrast, the huge entropies in Eq. (0.14) are associated with degrees-of-freedom that are maximally entangled with the outgoing radiation and therefore correspond to an effect of the "infalling partners" to the radiation. Thus, Eq. (0.14) represents a distinct (and much stronger) phenomenon imposed by locality.

Non-exotic atmosphere: Assumption 2.c is taken to be effectively equivalent to 1.b(i). Thus, we only suppose that the black hole does not contain an atmosphere of super-entropic exotic matter. From Eq. (0.1)

$$I(N:R) \leq \eta\, S_{\rm BH}, \qquad \text{with } \eta \ll 1, \tag{0.15}$$

and combining Eqs. (0.14) and (0.15) yields the contradiction

$$1 \leq \varepsilon + \eta \ll 1, \tag{0.16}$$

whatever the details of the radiation process. ■

Mixed state black holes

In this section we generalize our results to show that they apply even when the matter that collapsed to form the black hole is not pure. We start with a more general review of generic black hole radiation necessary to analyse such scenarios.

Generics of black hole radiation: In the main text we considered a black hole with (initial) thermodynamic entropy $S_{\rm BH}$ which can completely evaporate into a net pure state of radiation. As discussed, the generic evaporative dynamics of such a black hole may be captured by Levy's lemma for the random sampling of subsystems from an initially pure state consisting of $S_{\rm BH}$ qunats [3]. This either assumes the infallen matter is pure (as in the main text) or ignores it entirely. Throughout, we set Boltzmann's constant to unity and work with natural logarithms leading to the measure of qunats (i.e., $\ln 2$ times the number of qubits).

In order to extend our analysis to include infallen matter carrying some (von Neumann) entropy $S_{\rm matter}$, we need only take the initially pure state used above and replace it with a bipartite pure state consisting of two subsystems: $S_{\rm BH}$ qunats to represent the degrees-of-freedom that evaporate away as radiation; and a reference subsystem. Without loss of generality, the matter's entropy may be treated as entanglement between these two subsystems, however, here we shall simplify our analysis by assuming uniform entanglement between the black hole subsystem and $S_{\rm matter}$ reference qunats. The generic properties of the radiation may then again be studied by random sampling the former subsystem to simulate the production of radiation [3].

The behavior is generic and for our purposes may be summarized in terms of the radiation's von Neumann entropy, $S(R)$, as a function of the number of qunats in this radiation subsystem. One finds [3] that $S(R)$ initially increases by one qunat for every extra qunat in R, until it contains $\frac{1}{2}(S_{\rm BH}+S_{\rm matter})$ qunats. From that stage on it decreases by one qunat for every extra qunat in R until it drops to $S_{\rm matter}$ when R contains $S_{\rm BH}$ qunats and the black hole has completely evaporated.

Because the von Neumann entropy of a randomly selected subsystem only depends on the size of that subsystem, the same behavior is found whether R above represents the early or late epoch radiation with respect to any arbitrary split. Further, in the simplest case where we choose the joint radiation (R, R') to correspond to the net radiation from a completely evaporated black hole we may immediately write down the generic behavior for the quantum mutual information $I(R':R)$.

In particular, $I(R':R)$ starts from zero when R consists of zero qunats. From then on, it increases by two qunats for every extra qunat in R until $I(R':R)$ reaches $S_{\rm BH}-S_{\rm matter}$ when R contains $\frac{1}{2}(S_{\rm BH}-S_{\rm matter})$ qunats. From that stage on until R contains $\frac{1}{2}(S_{\rm BH}+S_{\rm matter})$ qunats $I(R':R)$ remains constant, after which $I(R':R)$ decreases by two qunats for every extra qunat in R until it drops to zero once R contains the full $S_{\rm BH}$ qunats of the completely evaporated black hole [3]. Interestingly, it is during the region where $I(R':R)$ is constant that the information about the infallen matter becomes encoded into R for the first time [3]. Finally, setting $S_{\rm matter}$ to zero gives the 'standard' behavior for $S(R)$ and $I(R':R)$ upon which the results in the main text are derived.

From the above, we are motivated to generalize the Page time: we define *any* time where $I(R':R)$ is maximal a (generalized) Page time; the earliest such time the 'initial Page time'; and the latest the 'final Page time'. Prior to the initial Page time, the quantum information about the initial infallen matter is encoded entirely within the black hole interior [3]. After the final Page time this information is encoded entirely within the radiation [3].

Including infallen matter: Let us start with a consideration of how the reasoning in Theorem 2 becomes modified by the presence of infallen matter carrying entropy.

Theorem 2 generalized: In the main text we did not explicitly include entropy associated with infallen matter. Fig. 4 shows the most general scenario. Subsystem I denotes the matter that falls into the region surrounding the black hole where radiation is produced. Thus, we suppose that late epoch radiation can in principle come from the joint subsystem (N, I). In this figure we also include subsystem $I_{\rm early}$ denoting matter that has fallen into the region surrounding the black

hole at an earlier epoch or indeed matter that may have collapsed to form the black hole in the first place.

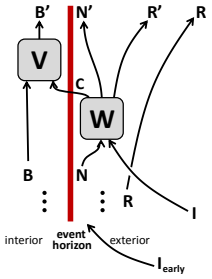

Figure 4. Quantum circuit diagram for evaporation of a quantum black hole with causal horizon and infallen matter. Subsystem I denotes infallen matter that falls into the region surrounding the black hole to participate in late epoch radiation generation. (This does not exclude the possibility that the matter falls directly into the black hole.) Subsystem I_{early} denotes matter infalling at earlier times or even that collapses to form the original black hole.

As in the main text we apply strong subadditivity:

$$I(R' : R) \leq I(C, N', R' : R) = I(N, I : R) = I(N : R). \tag{0.17}$$

Here, we used the fact that joint subsystems (C, N', R') and (N, I) are unitarily related. Finally, the most natural assumption is that the infallen matter I is independent of the quantum state of the black hole, (B, N), or its early epoch radiation R. The original inequality of Eq. (0.6) for pure-state black holes is thus found to still hold in the presence of mixed infallen matter.

From the summary above of generic radiation production including infallen matter we have enough to generalize Theorem 2. As in the main text, we take R to be all the early epoch radiation until the Page time (for this theorem we may take any generalized Page time), and we let R' denote all the radiation generated from the Page time onward until the black hole has shrunk to a size much smaller than the original black hole (say roughly $\varepsilon/2$ of its original area), but still much larger than the Planck scale. In this case, instead of Eq. (0.9), we have

$$I(R' : R) = (1 - \varepsilon)\, S_{\text{BH}} - S_{\text{matter}}, \qquad \varepsilon \ll 1, \tag{0.18}$$

where $S_{\text{matter}} \equiv S(I_{\text{early}}, I)$ is the net entropy in all the infallen matter. Combining this with Eqs. (0.15) and (0.17) gives

$$1 - \frac{S_{\text{matter}}}{S_{\text{BH}}} \leq \epsilon + \eta \ll 1. \tag{0.19}$$

Once again we obtain a contradiction except in the extreme case of a black hole whose net infallen matter contains virtually as much entropy as the entire black hole's original entropy S_{BH}.

Theorem 1 generalized: It is simple enough to repeat the above reasoning for Theorem 1, where we no longer make use of locality. In this case, we may still use Fig. 4 provided we ignore the no-communication decomposition structure. In particular, strong subadditivity yields

$$I(R' : R) \leq I(B', N', R' : R) = I(B, N, I : R) = I(B, N : R). \tag{0.20}$$

Here, we use the fact that joint subsystems (B', N', R') and (B, N, I) are unitarily related. Again, the most natural assumption is that the infallen matter I is independent of the quantum state of the black hole, (B, N), or its early epoch radiation R.

Applying Eq. (0.18) to any Page time then tells us that for a unitarily and completely evaporating black hole

$$I(B, N : R) \geq (1 - \varepsilon)\, S_{\text{BH}} - S_{\text{matter}}, \qquad \varepsilon \ll 1. \tag{0.21}$$

To simplify our argument, we shall suppose that the infallen matter (I_{early}, I) has actually entered the black hole. In that case, for any times prior to the initial Page time, the infallen matter's

external reference qunats are maximally entangled with some subsystem of the black hole interior [3]. We shall label the orthogonal complement of this subsystem within B as $B^{\perp}$. It is clear that: i) $(B^{\perp}, N, R)$ can be treated as a pure quantum state; and ii) $I(B, N{:}R) = I(B^{\perp}, N{:}R)$. So that

$$I(B^{\perp} : R) + I(N : R) \geq (1 - \varepsilon)\, S_{\rm BH} - S_{\rm matter}, \qquad \varepsilon \ll 1. \tag{0.22}$$

To ensure that our infalling observer is not passing through an atmosphere of exotic matter before they reach the horizon, Eq. (0.1) for a large black hole implies that Eq. (0.22) reduces to

$$I(B^{\perp} : R) \gtrsim S_{\rm BH} - S_{\rm matter}. \tag{0.23}$$

Since $\ln(\dim(B^{\perp})) = \ln(\dim(B)) - S_{\rm matter}$ by construction, we find the trivial bound

$$\ln(\dim(B)) \gtrsim \frac{1}{2}(S_{\rm BH} + S_{\rm matter}). \tag{0.24}$$

If this bound were saturated, then the huge thermodynamic entropy in B would imply that an infalling observer would immediately encounter an incredibly mixed state with correspondingly huge energies as soon as they passed the horizon. They would immediately encounter an 'energetic curtain' or firewall upon entering the black hole. To ensure, therefore that assumption 1.b holds, the above bound must be far from saturation, i.e.,

$$\ln(\dim(B)) \gg \frac{1}{2}(S_{\rm BH} + S_{\rm matter}), \tag{0.25}$$

where B is the black hole at the *initial* Page time.

However, assumption 1.c would require that the left- and right-hand-sides of Eq. (0.25) should be nearly equal. As with the generalization of Theorem 2, we again obtain a contradiction, in this case, however, apparently independent of the amount infallen matter.

References

1. S. W. Hawking, Phys. Rev. D **14**, 2460 (1976).
2. A. Almheiri, et al., J. High Energy Phys. **02** (2013) 062.
3. S. L. Braunstein, arXiv:0907.1190v1 (2009); published as S. L. Braunstein, S. Pirandola and K. Życzkowski, Phys. Rev. Lett. **110**, 101301 (2013);
4. D. Harlow and P. Hayden, arXiv:1301.4504.
5. D. A. Lowe, J. Polchinski, L. Susskind, L. Thorlacius, and J. Uglum, Phys. Rev. D **52**, 6997 (1995).
6. G. 't Hooft, in *Salamfestschrift: A Collection of Talks*, edited by A. Ali, J. Ellis, and S. Randjbar-Daemi (World Scientific, Singapore, 1993), Vol. 4, p. 284.
7. L. Susskind, J. Math. Phys. **36**, 6377 (1995).
8. J. M. Maldacena, JHEP **04**, 021 (2003).
9. V. P. Frolov and A. Zelnikov, Introduction to Black Hole Physics (OUP, Oxford, 2011), p. 409.
10. J. Eisert, M. Cramer and M. B. Plenio, Rev. Mod. Phys. **82**, 277 (2010).
11. S. D. H. Hsu and D. Reeb, Phys. Lett. B **658**, 244 (2008).
12. D. N. Page, Phys. Rev. Lett. **71**, 3743 (1993).
13. S. L. Braunstein and M. K. Patra Phys. Rev. Lett. **107**, 071302 (2011).
14. J. D. Bekenstein, Phys. Rev. D **49**, 1912 (1994).
15. S. B. Giddings, Phys. Rev. D **51**, 6860 (1995).
16. C. Rovelli, arXiv:1710.00218.
17. M. K. Parikh and F. Wilczek, Phys. Rev. Lett. **85**, 5042 (2000).
18. R. Bousso, JHEP **06**, 028 (1999).
19. H. Kawai and Y. Yokokura, Int. J. Mod. Phys. A **30**, 1550091 (2015).
20. A. Strominger, presented at the 1994 Les Houches Summer School "Fluctuating Geometries in Statistical Mechanics and Field Theory," arXiv:hep-th/9501071
21. T. Eggeling, D. Schilingemann and R. F. Werner, Europhys. Lett. **57**, 782 (2002).
22. S. Giddings, in *1994 Trieste Summer School in High Energy Physics and Cosmology* (World Scientific, NY, 1995).