Preprints.org

Article

Not peer-reviewed version

No Small Observer Can Verify a Black Hole Firewall

Michael Timothy Bennett *

Posted Date: 4 March 2026

doi: 10.20944/preprints202603.0227.v1

Keywords: black hole firewall; AMPS; Stack Theory

Preprints.org is a free multidisciplinary platform providing preprint service that is dedicated to making early versions of research outputs permanently available and citable. Preprints posted at Preprints.org appear in Web of Science, Crossref, Google Scholar, Scilit, Europe PMC.

Copyright: This open access article is published under a Creative Commons CC BY 4.0 license, which permit the free download, distribution, and reuse, provided that the author and preprint are cited in any reuse.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

Disclaimer/Publisher's Note: The statements, opinions, and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions, or products referred to in the content.

Article

No Small Observer Can Verify a Black Hole Firewall

Michael Timothy Bennett

Australian National University; michael.bennett@anu.edu.au

Abstract

Quantum gravity is often probed by thought experiments that assume unlimited classical control over quantum systems. The firewall paradox is one such thought experiment. It asks if one observer can decode early Hawking radiation, and then later compare with degrees of freedom behind the horizon. Here I show that this protocol is constrained by an instruction budget. If the correct decoding depends on the black hole microstate, a hidden internal configuration, then the observer must specify which decoder to use. The number of decoders needed to cover all microstates grows exponentially with the black hole entropy, so the decoder specification requires a number of bits that scales with that entropy. A covariant entropy bound limits how much information can be stored within a bounded causal patch. This enforces a horizon-scale lower bound on the region required for single-observer verification, even when computational complexity is ignored. More broadly, the result highlights a principle for quantum gravity experiments. Control information is physical, and bounds on storage can decide what is testable.

Keywords: black hole firewall; AMPS; Stack Theory

1. Introduction

Black hole complementarity is the idea that two observers can each use a consistent description, even if those descriptions cannot be combined into one single measurement record, because no single observer can access both sets of data [1,2]. One description has a smooth horizon for infallers. The other treats evaporation as unitary, meaning information is not destroyed and in principle can be reconstructed from the outgoing Hawking radiation.

I'll begin with a little background on these terms. A quantum state is a rule for predicting measurement outcomes. A *pure* state is the special case where that rule comes from one wavefunction[1]. A *mixed* state is what you get when you have ordinary uncertainty about the preparation, or when you look at only part of an entangled system. A purifier is an *extra* system on top of all this. A purifier together with the mixed system forms a pure joint state.

Meanwhile a black hole horizon is not a surface, but a point of no return where the gravity is too strong for anything to get back out[2]. If you fall freely in a gravitational field, you feel like you are floating in deep space, so it is believed that the horizon is "smooth" in the sense that a person passing through it will not detect any difference. Black hole evaporation is widely believed to be unitary. Unitary means information is preserved (transformed but never destroyed), at least in principle. Evaporation refers to Hawking radiation, the process by which a black hole radiates away its mass and energy. Unitarity then says the information about what fell in is encoded in correlations of the outgoing radiation rather than being destroyed.

So now you have some background, assume two people named Alice and Bob. Alice stays outside the black hole, and Bob falls in. Alice collects the Hawking radiation and confirms it is unitary, but never sees the inside. Bob passes through the horizon and confirms it is smooth, but can never send a

[1] Mathematically it can be written as a unit vector $|\psi\rangle$, or as a rank-one density matrix $\rho = |\psi\rangle\langle\psi|$.
[2] Since light can't get out, might look like a pitch black surface.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

message back. Subjectively, physics holds up for both of them, but objectively their realities contradict one another. Complementarity holds that one should discard the objective view because no single measurement can access both sides [1].

The firewall paradox, often called AMPS after Almheiri, Marolf, Polchinski and Sully, attacks this by proposing a single-observer test [3]. If evaporation is unitary, then the early radiation R should contain enough information to predict a late outgoing mode B. If the horizon is smooth, then B should also be maximally entangled with an interior partner A. Maximally entangled means A and B share the strongest possible quantum correlations; each subsystem alone looks completely random, but together they behave like a perfectly correlated pair. Quantum mechanics does not allow one system to be maximally entangled with two independent systems. Breaking the entanglement between A and B "snaps" the vacuum state[3], creating a firewall of high-energy particles that blasts anything crossing the horizon. But a firewall contradicts the smooth horizon. This is the monogamy problem: B cannot be entangled with *both* A and R.

AMPS proposed that one observer could collect the early radiation outside, distil the information from it, then jump into the black hole and compare with the interior mode. A single subjective observer would then witness the contradiction. To prevent this, the universe would have to "burn" them with a firewall before they cross the horizon, which removes the smooth horizon. This expresses a tension between quantum theory and gravity. The structural connection to observer-dependent paradoxes has also been noted through reframings as an extended Wigner's friend scenario [4].

However, if no single observer can perform the needed joint experiment, then the AMPS firewall paradox is *not* a problem. Instead of a paradox, there would be a constraint on which measurements are physically meaningful in quantum gravity.

1.1. What This Paper Asks

This paper asks whether the AMPS verification protocol can be implemented inside one bounded causal patch once I include the control information needed to choose the correct decoder. I model the decoder choice as physical control information, meaning the distinguishable control states that select which decoding procedure is applied [5,6]. Treating the covariant entropy bound as a memory ceiling limits how many distinct decoder choices any bounded patch can hold [7]. Separately, the alpha-bit analysis of Hayden and Penington bounds how large a code subspace can share one fixed, state-independent decoder at a given accuracy, even for mixtures and for states entangled with an external reference system [8]. In evaporating setups this tension appears as state dependence of interior reconstructions for mixed states [9,10].

The implication is that some firewall contradictions may be untestable by any observer who only accesses a subhorizon patch. That shifts the target for quantum gravity. It is not only about what *exists*, but what can be *certified*.

1.2. The Protocol

Hawking radiation is the outgoing quantum radiation emitted by a black hole as it evaporates. I label the early Hawking radiation as R. It is the radiation already emitted to infinity. An old black hole is one that has emitted enough Hawking radiation that R can in principle predict properties of the late radiation. I label a late outgoing mode as B. This is one of the later quanta emitted after the black hole is old. I label the would-be interior partner of B as A. If the horizon is smooth, then A and B should be strongly entangled.

A decoder is a physical procedure, a map or circuit, that acts on R and outputs a system that purifies B. Purifies means that the output and B together form a pure state, meaning the joint system can be described by one wavefunction rather than a probabilistic mixture. A causal patch is the

[3] In quantum field theory, "empty space" is not nothing. It is a pattern of entanglement between adjacent regions. In the Rindler and Unruh picture this entanglement is what makes the vacuum look thermal to accelerated observers. Breaking this pattern releases energy.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

region of spacetime whose degrees of freedom can still influence the infaller before the comparison is completed.

The AMPS-style verification task can be stated as five steps.

1. Collect the early radiation R outside the horizon.
2. Wait until the black hole emits a fresh outgoing mode B.
3. Choose a decoder and apply it to R to produce a candidate purifier of B.
4. Cross the horizon and access the interior partner mode A of the same B.
5. Compare the candidate purifier with A and B inside one causal patch.

The first three steps happen outside. The last two steps happen inside. The word verify means a single observer can do all five steps using only the degrees of freedom that fit in their patch.

1.3. Where The Obstruction

If one fixed decoder worked uniformly for all black hole microstates, the story would be simple. You would bring that one decoder and run it. Holographic reconstruction bounds that demand uniformity on mixtures and on states entangled with an external reference system imply this uniformity fails in general. Instead, one fixed decoder can only be trusted on a limited code subspace. Different microstates require different decoders.

This paper just points out the obvious. The decoder choice itself is physical. At verification time the infaller must carry a control register K that selects which decoder is being applied. If there are many decoders, then K must have many reliably distinguishable internal states. The covariant entropy bound acts as a ceiling on how many such internal states can fit in a bounded patch.

1.4. Why Waiting Longer Doesn't Help

It is tempting to think the protocol could be saved by taking more time. In Stack Theory, the operational notion of time at a layer is the count of changes that the layer can represent [6]. If nothing that your vocabulary can express changes, then no time passes at that layer. This means many step descriptions can be rephrased as bigger state descriptions by storing intermediate results [11].

The firewall verification task is different. It is a co-instantiation task. Co-instantiation means all required ingredients are true at the same objective instant for one observer. The observer needs one moment in which the late mode B, its interior partner A, the decoded purifier, and the decoder selector K are all physically present inside one causal patch. Seeing each ingredient at some point along a long worldline does not create one joint measurement record. Section S3 of the supplement formalises this gap between ingredient-wise occurrence and co-instantiation.

This is not the Harlow–Hayden time barrier, which is a claim about how long the distillation computation takes [12]. Even if an oracle, meaning a hypothetical device that performs computation for free, performed the computation instantly, the physical control state that selects the right decoding operation still has to fit in the patch [13,14].

1.5. The Theorem

Theorem 1 turns the above story into one inequality. It uses three numbers.

S_{BH} is the black hole horizon entropy in bits. It is $\log_2 |\Phi_{\mathrm{micro}}|$ where Φ_{micro} is the set of mutually exclusive microstates compatible with the same exterior. ρ^2 is the fraction of the horizon area that bounds the infaller patch. α_* is a number between 0 and 1 that controls how large a code subspace can share one fixed decoder at fixed error tolerance. It is defined precisely in Box 3.

The theorem shows that a verifier confined to a patch with boundary area fraction ρ^2 can only succeed on all microstates if

$$\rho^2 \geq 1 - \alpha_* - \frac{c_\varepsilon}{S_{\mathrm{BH}}}.$$

For large black holes the correction term is negligible, so the punchline is $\rho^2 \gtrsim 1 - \alpha_*$. In a representative case where one decoder works uniformly on a code subspace carrying 20% of the black hole

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

microstate entropy, the decoder label alone costs about 80% of the entropy in control bits and demands access to about 80% of the horizon area.

Figure 1 plots the large entropy form of the theorem. It tells you the minimum patch size fraction $\rho_{\min}$ required for each reconstruction fraction α_*. The curve is $\rho_{\min} = \sqrt{1-\alpha_*}$ when subleading constants are neglected.

The Extended Data figures in the supplement are toy computations. They do not simulate black holes. They illustrate the two counting steps: how join capacity can exceed a patch ceiling, and how decoder coverage scales with a code subspace size bound.

If you want to reproduce these toy plots, see Section 7. The exact toy parameters are listed there and are also recorded in the CSV source data files. For a plain English description of each figure and its source data, see `figures/README.md` and `source_data/README.md`. One warning helps avoid confusion. The toy exponent parameter called α in Extended Data Figure 2 is not the same object as the theorem parameter α_*. The theorem parameter α_* is the geometric reconstruction fraction defined in Box 3. The toy parameter α is just a dial that sets a subset size $d_{\max} = \lfloor 2^{\alpha n} \rfloor$ in a finite set cover model.

1.6. Contributions

Many firewall discussions focus on computation time, asking whether the distillation circuit is too slow to run before the infaller reaches the singularity, the region deep inside where curvature becomes extreme and semiclassical descriptions break down. This paper isolates a different bottleneck: the physical memory needed to specify which distillation circuit to run when the correct circuit depends on hidden microscopic details of the black hole state. In this model, the instructions for the right reconstruction can be the scarce resource before the data.

Concretely, I make the following contributions. First, I treat the decoder label as a physical control state that must fit inside the same causal patch as the final comparison. Second, using alpha-bit reconstruction bounds, I translate state dependence into a lower bound on how many distinct decoders are needed to cover all microstates at fixed error tolerance. Third, I translate that decoder count into a lower bound on control bits, using the Stack Theory capacity definition in Box 2. Fourth, I combine that control bit requirement with the covariant entropy bound, treated as a patch memory ceiling, to obtain an explicit lower bound on the required horizon area fraction in Theorem 1. Fifth, I include a deterministic toy code script, source data, and plots that illustrate the counting steps without requiring any quantum gravity simulation.

The principle that control information is a physical resource in holographic settings is shared with Kubicki, May and Pérez-García [15]. They apply it to general bulk computation. The present paper applies this lens specifically to AMPS verification and combines it with alpha-bit coverage bounds to produce the horizon area fraction inequality in Theorem 1.

1.7. Outline

If you want the formal definitions, go to Section 2. If you want the main result, go to Theorem 1 and Figure 1. If you want to reproduce the plots, see Section 7.

2. Glossary and Notation

This paper aims to be readable by physicists, computer scientists, and information theorists. It uses a small vocabulary from quantum gravity and from quantum information. I define those terms here and then stick to them. The main result is Theorem 1. If you want to reproduce the figures, see Section 7.

2.1. Information Capacity

I use a simple capacity measure that counts how many distinct internal states a physical system can reliably represent. I write $\mathsf{I}(v)$ for the number of bits of distinction available to a vocabulary v. Box 2 gives the formal Stack Theory definitions. In plain English, if a patch can only support 2^{10}

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

reliably distinct internal states, then it has at most 10 bits of usable capacity no matter how clever the computation is.

2.2. Time as Change

Stack Theory treats time at a vocabulary as the count of changes that the vocabulary can represent [6]. Given an underlying trajectory of world states, the induced layer trajectory deletes repeated frames that look identical at that vocabulary. If the encoding stops changing, the layer clock stops too.

If nothing you can notice changes, then no time passes for you at that descriptive level. This matters because waiting longer is only helpful if it creates new distinguishable states. It cannot increase the number of distinct states that fit inside one causal patch at one moment. The firewall verification task is a one-moment comparison, so it is governed by co-instantiation capacity rather than by elapsed time.

2.3. Basic Mathematical Conventions

Box 1: Basic concepts for new readers.

Cardinality. For a finite set X, the notation $|X|$ means the number of elements in X—the count of distinct items in the set.

Subset and power set. For sets X and Y, the notation $X \subseteq Y$ means every element of X is also an element of Y. A subset is just a smaller collection chosen from a larger list. The power set 2^Y is the set of all subsets of Y.

Union and intersection. For sets X and Y, the union $X \cup Y$ is the set of elements that are in X or in Y (combine). The intersection $X \cap Y$ is the set of elements that are in both X and Y (keep only the overlap).

Injective map. A function $f : X \to Y$ is injective if $f(x) = f(x')$ implies $x = x'$. Different inputs always give different outputs, so the input can be recovered uniquely from the output.

Floor and ceiling. For a real number x, the floor $\lfloor x \rfloor$ is the greatest integer that is not larger than x. The ceiling $\lceil x \rceil$ is the smallest integer that is not smaller than x. I use these when turning idealised real-valued bounds like $2^{\alpha n}$ into integer counts in the toy computations.

Orthonormal basis. An orthonormal basis is a list of mutually perpendicular unit vectors. In quantum mechanics it is the formal way to say a set of perfectly distinguishable pure states. When I write a microstate basis $\{|i\rangle\}$, I mean such a set of mutually distinguishable microstates.

Bits and base-two logarithms.

All logarithms are base two, so information is measured in bits. I use $\log_2$ so that one bit corresponds to a factor of two in the number of distinguishable states. If a physical system can be in N reliably distinguishable states, then it can carry at most $\log_2 N$ bits of information.

Fundamental constants.

I write c for the speed of light, G for Newton's gravitational constant, and $\hbar$ for the reduced Planck constant. They are shown explicitly only to make dimensions clear. If you prefer natural units you can set $c = G = \hbar = 1$ everywhere without changing any dimensionless inequality in this paper.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

Reliably distinguishable states.

I say two internal states are reliably distinguishable if there exists some measurement that can tell which one you have with a small error probability. This is the operational notion behind capacity bounds like the covariant entropy bound.

Operationally distinguishable states and procedures.

I say two states or procedures are operationally distinguishable if there exists some experiment that can tell them apart. For states, this means there is some measurement whose outcome statistics differ for the two states. For procedures, such as two decoder channels, this means there is some input state and some measurement on the output that produce different statistics depending on which procedure was applied. In plain English, if no possible test can notice a difference, then the two procedures are the same for all physical purposes.

In this paper I adopt the weakest possible counting convention. I only distinguish decoders to the extent that they differ on the input domain that the verifier can actually supply, at the tolerated error level. If two decoding procedures act identically on every protocol-relevant input state, then I treat them as the same decoder and they do not require two different control settings. This makes the decoder family size $|\mathcal{D}|$ as small as it can be, which makes the no-go theorem harder to prove.

Order one fractions.

When I say an order-one fraction of the horizon area, I mean a constant fraction like 20% or 80% that does not shrink as the entropy grows.

Box 2: Stack Theory primitives and a derived capacity measure.

An environment is a nonempty set Φ of mutually exclusive states. A program is any subset $p \subseteq \Phi$. A vocabulary is a set of programs $v \subseteq 2^{\Phi}$. A statement is a set of programs $l \subseteq v$ with nonempty truth set

$$T(l) := \bigcap_{p \in l} p, \qquad T(\varnothing) := \Phi.$$

The induced encoding is $\mathrm{Enc}_v(\phi) := \{p \in v : \phi \in p\}$ and the realised encoding set is $\mathrm{Im}(\mathrm{Enc}_v) := \{\mathrm{Enc}_v(\phi) : \phi \in \Phi\}$. The encoding capacity is $\mathsf{I}(v) := \log_2 |\mathrm{Im}(\mathrm{Enc}_v)|$.
A bounded physical region $\mathcal{R}$ admits at most $N_{\max}(\mathcal{R})$ reliably distinguishable internal states, so it has ceiling $\mathsf{I}_{\max}(\mathcal{R}) := \log_2 N_{\max}(\mathcal{R})$ bits. Any joint protocol that needs a vocabulary v with $\mathsf{I}(v) > \mathsf{I}_{\max}(\mathcal{R})$ cannot be instantiated inside one patch.

An environment Φ is the set of mutually exclusive ways the world could be. It can be infinite. A program $p \subseteq \Phi$ is one yes or no question about the world. The answer is yes exactly on the states in p. A vocabulary v is the list of programs the observer can actually implement. In this paper it is finite (in related work this is justified using the Bekenstein bound [16]).

A statement $l \subseteq v$ is a set of programs that can all be true at once. Its truth set $T(l)$ is the set of world states where every program in l answers yes. The layer L_v is the list of statements that are satisfiable, meaning they have a nonempty truth set.

The encoding $\mathrm{Enc}_v(\phi)$ is the list of programs in v that answer yes on the state ϕ. Two different world states are distinguishable by v only if they induce different encodings. The realised encoding set $\mathrm{Im}(\mathrm{Enc}_v)$ is the set of distinct encodings that actually occur as ϕ ranges over Φ. The encoding capacity is $\mathsf{I}(v) = \log_2 \big|\mathrm{Im}(\mathrm{Enc}_v)\big|$. It is the number of bits of distinction that v can represent about the world.

Even if Φ is infinite, a finite vocabulary can realise only finitely many distinct encodings. In fact $\big|\mathrm{Im}(\mathrm{Enc}_v)\big| \le 2^{|v|}$, so $\mathsf{I}(v) \le |v|$. So a finite list of questions carves an unbounded world into finitely many operational cases.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

A bounded patch $\mathcal{R}$ can only hold finitely many reliably distinguishable internal states. I write $\mathsf{I}_{\max}(\mathcal{R})$ for the corresponding bit ceiling. If a protocol needs more distinctions than that ceiling, no amount of extra computation can make it fit in one causal patch.

A supposed paradox becomes operational only if the relevant joint vocabulary fits inside one causal patch. If it does not, the paradox is a request for an unphysical measurement.

For the main theorem, the only quantity I will use from Box 2 is the key capacity $\mathsf{I}(v_K)$. It equals $\log_2$ of the number of reliably distinguishable control states of the decoder selector register. The full vocabulary structure is used in Supplementary Section S2 and in Extended Data Figure 1 where I discuss capacity complementarity from joins.

2.4. Toy Example

Take $\Phi = \{a, b, c, d\}$. Let $p_1 = \{a, b\}$ and $p_2 = \{a, c\}$. Let $v = \{p_1, p_2\}$. The encoding of each state is

World state ϕ	Encoding $\mathrm{Enc}_v(\phi)$
a	$\{p_1, p_2\}$
b	$\{p_1\}$
c	$\{p_2\}$
d	$\emptyset$

So the realised encoding set is $\mathrm{Im}(\mathrm{Enc}_v) = \{\{p_1, p_2\}, \{p_1\}, \{p_2\}, \emptyset\}$. It has $\left|\mathrm{Im}(\mathrm{Enc}_v)\right| = 4$ elements, so $\mathsf{I}(v) = \log_2 4 = 2$ bits.

Now split v into two smaller vocabularies $v_1 = \{p_1\}$ and $v_2 = \{p_2\}$. Each one has only two realised encodings, so $\mathsf{I}(v_1) = \mathsf{I}(v_2) = 1$ bit. But their join is $v_1 \vee v_2 = \{p_1, p_2\}$, which is the original v, and so $\mathsf{I}(v_1 \vee v_2) = 2$ bits. If a patch had a ceiling between 1 and 2 bits, then each description would fit on its own, but the joint description would not fit. This is the finite toy analogue of complementarity as a join overflow. Extended Data Figure 1 repeats this pattern at larger scale using random programs.

3. Single Patch Versus Distributed

AMPS refers to the firewall argument of Almheiri, Marolf, Polchinski and Sully [3]. It is often told using a distillation step outside the horizon and a comparison step inside. That is a distributed protocol. The firewall contradiction is instead a statement about what one observer can certify. The observer is supposed to know that B is purified by R and also know that B is entangled with an interior partner A.

I therefore distinguish two notions.

Definition 1 (Single patch verifier)**.** *A single patch verifier is a protocol whose decoding choice and verification outcome are both implemented by degrees of freedom confined to one infaller causal patch at the moment of verification. External preprocessing is allowed only insofar as its output is carried into the patch as physical state.*

A protocol counts as single patch only if the choice of decoder and the final yes or no verdict are both carried by physical degrees of freedom inside the same infaller patch at the moment of comparison. If some outside agent does work and then merely tells the infaller which decoder to use, that instruction still has to be stored inside the patch as a physical control state. If it does not fit, the protocol is distributed rather than single patch.

Remark 1. *If you allow a distributed protocol, then you have already granted complementarity. You have accepted that no single patch contains the joint refinement of the exterior and interior descriptions.*

The next section proves a no-go theorem for single patch verifiers in a controlled reconstruction model.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

4. No-go Theorem From Control and Reconstruction Limits

Box 3: Assumptions for the no-go theorem.

1. Patch capacity ceiling. A causal patch $\mathcal{R}$ with maximal boundary cross-sectional area $A(\mathcal{R})$ supports at most $2^{A(\mathcal{R})/(4\ell_p^2 \ln 2)}$ reliably distinguishable internal states. This treats the covariant entropy bound as a capacity ceiling [7]. This is an interpretation step. The covariant entropy bound is formally a bound on entropy on a light sheet, not a direct statement about the Hilbert space dimension of a spatial region. Here I adopt the working assumption that it upper bounds the number of reliably distinguishable internal register states that can be stored inside the patch. This is motivated by holographic state counting in semiclassical settings and by the Bekenstein bound for bounded systems [17]. The theorem below is conditional on this memory ceiling interpretation.
2. Microstate coverage requirement. Fix a set Φ_{micro} of mutually exclusive black hole microcases that are compatible with the same exterior semiclassical data. I do not assume the full Stack Theory environment Φ is finite. The only finiteness input is that this microcase list is finite and has size $|\Phi_{\text{micro}}| = 2^{S_{\text{BH}}}$. You can view this as a statement about an induced layer, which in Stack Theory typically represents an embodied organism embedded in an environment [18,19]. There exists some finite micro vocabulary whose realised encodings label these microcases and whose encoding capacity is S_{BH} bits. If you prefer the usual quantum description, you can take Φ_{micro} to be an orthonormal microstate basis of a Hilbert space $\mathcal{H}_{\text{micro}}$ with $\dim \mathcal{H}_{\text{micro}} = |\Phi_{\text{micro}}|$. The verifier is required to succeed for every microstate $\phi \in \Phi_{\text{micro}}$. I assume the verifier does not have prior knowledge of which microstate is realised. If the microstate were known, a single decoder would suffice and no control key would be needed. When I say a decoder succeeds on a subset $\Phi'_{\text{micro}} \subseteq \Phi_{\text{micro}}$, I mean it succeeds uniformly on all states supported on the code subspace spanned by Φ'_{micro}. This includes mixtures and states entangled with an external reference. For the counting argument I associate each decoder D with one such success set $\Phi^{(D)}_{\text{micro}} \subseteq \Phi_{\text{micro}}$. I don't use the typical Stack Theory task notation here because I'm trying to reduce the overhead in learning background Stack Theory definitions, and the learning and inference definitions Stack Theory normally uses are not strictly necessary for this paper. However just to link what I'm talking about here back to a normal Stack Theory interpretations, a decoder can be understood as an embodied policy. Success in decoding is success in a task. A causal patch is an embodied system embedded in its environment.
3. State-dependence bound from holographic reconstruction. In a holographic reconstruction setting, and in particular in evaporating black hole plus bath models, any fixed state-independent reconstruction acting on the verifier-accessible region can succeed uniformly on at most $2^{\alpha_* S_{\text{BH}} + c_\varepsilon}$ microstates at error tolerance ε, where $\alpha_* = (A_2 - A_1)/A_0$ and c_ε does not scale with S_{BH} [8–10]. The parameter α_* lies between zero and one. If α_* is not close to one then one decoder covers only a tiny exponential fraction of microstates.
4. Distinguishable control states. Operationally distinct decoders for the verification task require operationally distinguishable internal control states for the selector register.

The assumptions are listed in Box 3. From assumption 1 (patch capacity cieling) I have a hard memory ceiling for anything that must sit inside one causal patch. From assumption 2 (the microstate coverage requirement) I have a well defined requirement for successful decoding. The verifier must work for every microstate in Φ_{micro}, not just for a lucky subset. When I say a decoder works on a subset of microstates I mean it works uniformly on every state in the code subspace they span, including mixed and reference-entangled states. Assumption 3 is the holographic input. It says that if the reconstruction is state independent, one fixed decoder can only cover an exponentially small code

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

subspace of microstates unless the geometric fraction parameter α_* is close to one. Assumption 4 is the operational bridge. Different decoders require different physical control settings. I count decoders in the most forgiving way. Only differences that matter on protocol-relevant inputs are counted as different control settings. The theorem below is what you get when you combine these ceilings and then ask one patch to hold the decoder choice.

4.1. A Covariant Capacity Ceiling for a Subhorizon Patch

For a nonrotating, uncharged black hole of mass M, the horizon radius is the Schwarzschild radius $R_s := 2GM/c^2$ and the horizon area is $A_0 := 4\pi R_s^2$. Let $\ell_p^2 := \hbar G/c^3$ be the Planck area. The horizon entropy in bits is given by [20,21]

$$S_{\mathrm{BH}} := \frac{A_0}{4\ell_p^2 \ln 2}. \tag{1}$$

The factor $\ln 2$ converts the usual natural logarithm convention into base two so the unit is bits.

I model the infaller as confined to a bounded causal patch $\mathcal{R}$. Let $A(\mathcal{R})$ denote the maximal area of a codimension-two cross-section of the boundary of $\mathcal{R}$. I adopt Assumption 1. I treat the covariant entropy bound as a patch capacity ceiling. In bits [7],

$$\mathsf{I}_{\max}(\mathcal{R}) \le \frac{A(\mathcal{R})}{4\ell_p^2 \ln 2}. \tag{2}$$

Suppose the patch boundary area satisfies $A(\mathcal{R}) \le \rho^2 A_0$ for some $\rho \in (0,1]$. This defines ρ^2 as the accessible area fraction of the horizon.

Lemma 1 (Covariant subhorizon ceiling)**.** *For any causal patch $\mathcal{R}$ with $A(\mathcal{R}) \le \rho^2 A_0$,*

$$\mathsf{I}_{\max}(\mathcal{R}) \le \rho^2 S_{\mathrm{BH}}.$$

Proof. Combine (2) with $A(\mathcal{R}) \le \rho^2 A_0$ and the definition (1). □

The inequality above counts only the control information needed to select a decoder. It does not budget for storing the early radiation data, the distilled output, or the final comparison record. Any such storage competes for the same finite patch capacity, so including those costs would only make the no-go stronger.

Note also that Lemma 1 avoids any appeal to weak-gravity assumptions about the verifier. It is an area-based ceiling for any information that must fit inside one infaller patch.

4.2. Decoder Means Physical Key

The AMPS verifier must select a decoder that distils a purifier of a late mode B from the early radiation R. Let $\mathcal{D}$ be the set of decoder channels that are distinct for this verification task. Two candidate procedures count as the same decoder if they are operationally indistinguishable on every protocol-relevant input state at the tolerated error level. This convention minimises $|\mathcal{D}|$ and therefore makes the key lower bound as weak as it can be.

It also means the argument is not about where the decoder description sits. The key can be explicit, like a written label. It can be implicit, like the internal configuration of a programmable device after it has interacted with R. At the moment the device applies the decoding map, it must be in one of $|\mathcal{D}|$ operationally distinguishable configurations. That is the physical key.

Selecting among $|\mathcal{D}|$ operationally distinct decoders is a control problem that requires physical state inside the patch. I formalise this as control information in the Stack Theory sense [5,6]. A selector register that can implement any $D \in \mathcal{D}$ needs at least $|\mathcal{D}|$ reliably distinguishable settings. So its key vocabulary must satisfy $\mathsf{I}(v_K) \ge \log_2 |\mathcal{D}|$ by the capacity definition in Box 2. The next subsection

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

turns that informal point into a single counting inequality that also accounts for limited reconstruction coverage.

Quantum programming does not evade the key.

In this paragraph the word program is used in the programmable quantum processor sense. It refers to a control register state that selects an operation. This is not the same as a Stack Theory program, which is a subset of an environment. If one models the selector as a quantum program register, the conclusion is strengthened. For exact deterministic programming of a finite set of distinct unitaries, the program states must be mutually orthogonal, so the program register dimension is at least the number of implemented unitaries [13]. Approximate programming relaxes orthogonality, but quantitative results still impose nontrivial resource costs as the target family grows and the tolerated error shrinks. Kubicki, Palazuelos and Pérez-García provide quantitative lower bounds for approximate programmable devices. These bounds show that the required program register dimension grows at least polynomially in the target system dimension and in $1/\varepsilon$ for natural families of measurements and channels. I use this only to support the qualitative point that allowing a small error does not make the physical program register free [14]. The inequality in Theorem 1 uses exact distinguishability of the control states. In the approximate regime the required program register dimension can exceed $|\mathcal{D}|$, which would only strengthen the bound. A related use of programmable processor bounds together with gravitational entropy bounds to constrain bulk computation in AdS/CFT appears in Kubicki, May and Pérez-García [15]. My argument targets firewall verification and combines this programming requirement with alpha-bit coverage bounds to obtain an explicit horizon area fraction requirement.

The decoder key bound below is the step that closes the common loophole. A short classical description of a decoder is not a physical key unless that description is instantiated inside the patch as a distinguishable control state.

4.3. Reconstruction Bounds Limit Coverage

I now use a mainstream model class in which the state dependence of interior reconstruction is controlled. Concretely, I work in evaporating black hole models where the black hole is coupled to a non-gravitating bath [9,10]. The bath is an ordinary quantum system that stores Hawking radiation and can be measured without gravitational backreaction. The early radiation lives in bath degrees of freedom that I call R. The remaining black hole is described by the gravitational region or, in AdS/CFT language, by a boundary CFT. Interior reconstruction is described by entanglement wedge reconstruction. In plain English, this means that some interior operators can be represented as operators acting on exterior degrees of freedom, but only on a limited code subspace of microstates. The size of that code subspace is controlled by a geometric quantity that appears in Proposition 1.

Definition 2 (Error metric)**.** *For density matrices ρ and σ on the same Hilbert space, define the trace distance*

$$\Delta(\rho,\sigma) := \tfrac{1}{2}\|\rho-\sigma\|_1.$$

Here $\|X\|_1 := \mathrm{Tr}\sqrt{X^\dagger X}$ is the trace norm. An error tolerance ε means I accept trace distance at most ε between the decoded state and the target state. Throughout, an "error at most ε" means the same thing.

The trace distance is a standard way to say how distinguishable two quantum states are. It is zero only when the states are identical. It is at most one. If the trace distance is at most ε, then no measurement can tell the two states apart with advantage larger than about ε.

Definition 3 (Decoder microstate coverage)**.** *Fix an error tolerance $\varepsilon \in (0,1)$. Fix the microstate set Φ_{micro} from Box 3. For a decoder channel $D \in \mathcal{D}$, I define $d_\varepsilon(D)$ to be the* maximum cardinality of a subset *$\Phi'_{\mathrm{micro}} \subseteq \Phi_{\mathrm{micro}}$ such that D reconstructs the purifier of B with error at most ε uniformly on all states supported*

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

on the code subspace spanned by Φ'_{micro}[4]*. Uniformly includes mixtures of those microstates and states entangled with an external reference. Let*

$$d_\varepsilon^{\max} := \max_{D \in \mathcal{D}} d_\varepsilon(D).$$

Think of $d_\varepsilon(D)$ as the number of mutually exclusive microstate cases that one fixed decoder D can handle[5]. Uniformly means the decoder works not just for each microstate case but also for arbitrary mixtures and even when the microstate is entangled with an external reference system. This is the regime relevant for the holographic reconstruction bounds I use later. The geometric parameter α_* below comes from a competition between two quantum *extremal* surfaces.

A quantum extremal surface is a surface that extremizes the generalized entropy in semiclassical gravity. The area gap $A_2 - A_1$ acts like an entanglement budget that controls how large a code subspace can share one fixed decoder.

Proposition 1 (Hayden–Penington microstate coverage ceiling)**.** *Work in the holographic reconstruction setting analysed by Hayden and Penington [8]. Let A be the boundary region available to the verifier. Let A_1 and A_2 be the areas of the two competing quantum extremal surfaces anchored to ∂A, with $A_2 > A_1$. Let A_0 be the horizon area so that $S_{\text{BH}} = A_0/(4\ell_p^2 \ln 2)$. Define*

$$\alpha_* := \frac{A_2 - A_1}{A_0} \in (0,1).$$

Fix an error tolerance $\varepsilon \in (0,1)$ measured in trace distance. There exists a constant c_ε that does not scale with S_{BH} such that the following holds.

Consider any single, state-independent reconstruction channel D acting on the verifier-accessible region A. Then

$$d_\varepsilon(D) \le 2^{\alpha_* S_{\text{BH}} + c_\varepsilon} \qquad \text{and hence} \qquad d_\varepsilon^{\max} \le 2^{\alpha_* S_{\text{BH}} + c_\varepsilon}.$$

Proof sketch. Hayden and Penington analyse entanglement wedge reconstruction for a boundary region A in a black hole code space entangled with an external reference system R [8]. Two quantum extremal surfaces with areas A_1 and A_2 compete. In their derivation, decoding the intermediate region between the two surfaces from A for all states in a chosen subspace reduces, at leading order in G_N, to requiring that the entropy $S(R)$ never exceed the area gap. More precisely, their decoding condition is equivalent to the condition discussed around their equation (4.12) in Section 4.

$$4G_N S(R)_\psi < A_2 - A_1$$

uniformly over those states ψ. Here G_N is Newton's constant. Up to unit conventions, an area term divided by $4G_N$ is the Bekenstein–Hawking entropy in natural logarithm units. In this paper I express S_{BH} in bits via $A_0/(4\ell_p^2 \ln 2)$, so the scaling matches after dividing by $\ln 2$. If R purifies a maximally mixed state on a d-dimensional subspace, then $S(R) \approx \ln d$. Requiring uniform success on all mixed and reference-entangled states therefore forces

$$\ln d \le \frac{A_2 - A_1}{4G_N} + O(1).$$

Hayden and Penington use natural logarithms. Dividing by $\ln 2$ converts to bits and yields $\log_2 d \le \alpha_* S_{\text{BH}} + O(1)$. In my notation I can take d to be $d_\varepsilon(D)$ from Definition 3, since any covered set of $d_\varepsilon(D)$

[4] In Stack Theory terms $d_\varepsilon(D)$ can be understood as weakness of the weakest correct policy, but in raw microstates rather than a higher level embodied language.

[5] For those familiar with Stack Theory, the number of mutually exclusive microstate cases a decoder can handle would be proportional to Stack Theory's weakness measure, if one interprets the decoder as an embodied policy as in [16].

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

orthonormal microstates spans a $d_\varepsilon(D)$-dimensional code subspace on which D works uniformly. I absorb fixed accuracy and finite-size corrections into the additive constant c_ε. □

Lemma 2 (Decoder keys are hidden entropy). *Let $\mathcal{M}$ be a* finite *microstate*[6] *set with $|\mathcal{M}| = 2^S$. Let $\mathcal{D}$ be a finite family of operationally distinct decoders.*

For each $D \in \mathcal{D}$, let $\mathcal{M}^{(D)} \subseteq \mathcal{M}$ be the set of microstates on which D succeeds uniformly at some fixed accuracy. Assume a uniform coverage bound

$$|\mathcal{M}^{(D)}| \le d_{\max} \qquad \text{for every } D \in \mathcal{D}.$$

If the protocol must succeed for every microstate, meaning $\bigcup_{D\in\mathcal{D}} \mathcal{M}^{(D)} = \mathcal{M}$, then

$$|\mathcal{D}| \ge \frac{|\mathcal{M}|}{d_{\max}} = \frac{2^S}{d_{\max}}.$$

Therefore any control register that must select among $\mathcal{D}$ at verification time must support at least $|\mathcal{D}|$ reliably distinguishable control states. Equivalently any key vocabulary v_K representing that register must satisfy

$$\mathsf{I}(v_K) \ge \log_2 |\mathcal{D}| \ge S - \log_2 d_{\max}.$$

Proof. The covering condition gives

$$|\mathcal{M}| \le \sum_{D\in\mathcal{D}} |\mathcal{M}^{(D)}| \le |\mathcal{D}|\, d_{\max}.$$

Rearrange to get the lower bound on $|\mathcal{D}|$. A selector register that can implement any $D \in \mathcal{D}$ needs at least one reliably distinguishable control state per operationally distinct decoder. So the register has at least $|\mathcal{D}|$ internal states and any representing vocabulary satisfies $\mathsf{I}(v_K) \ge \log_2 |\mathcal{D}|$. □

Interpretation. If each decoder label leaves at most $d_{\max}$ microstates unresolved, then the label must carry the rest of the micro information. So the instruction budget is a hidden entropy term.

4.4. No Subhorizon Firewall Verifier Theorem

I can now state the main no-go theorem. It is a single inequality.

Theorem 1 (No subhorizon AMPS verifier under reconstruction limits). *Assume an old black hole with horizon entropy S_{BH} in bits. Fix the microstate set Φ_{micro} from Box 3 and assume $|\Phi_{\mathrm{micro}}| = 2^{S_{\mathrm{BH}}}$. Fix an error tolerance $\varepsilon \in (0,1)$.*

Consider a single patch verifier whose entire implementation at verification time is confined to a causal patch $\mathcal{R}$ with $A(\mathcal{R}) \le \rho^2 A_0$ for some fixed $\rho \in (0,1]$. If the verifier must succeed for every microstate $\phi \in \Phi_{\mathrm{micro}}$ with error at most ε, then any such protocol requires a key vocabulary with

$$\mathsf{I}(v_K) \ge S_{\mathrm{BH}} - \log_2 d_\varepsilon^{\max}.$$

If Proposition 1 holds, then

$$\mathsf{I}(v_K) \ge (1-\alpha_*) S_{\mathrm{BH}} - c_\varepsilon.$$

Since the key must be physically instantiated inside $\mathcal{R}$, success requires

$$\mathsf{I}_{\max}(\mathcal{R}) \ge \mathsf{I}(v_K).$$

By Lemma 1, this is possible only if

$$\rho^2 S_{\mathrm{BH}} \ge S_{\mathrm{BH}} - \log_2 d_\varepsilon^{\max}.$$

[6] $\mathcal{M}$ is finite meaning it is not a complete description of the universe microstate like $\phi \in \Phi$ would be.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

Equivalently

$$\rho^2 \geq 1 - \frac{\log_2 d_\varepsilon^{\max}}{S_{\mathrm{BH}}}.$$

In particular, inserting Proposition 1 yields the necessary condition

$$\rho^2 \geq 1 - \alpha_* - \frac{c_\varepsilon}{S_{\mathrm{BH}}}.$$

For large S_{BH}, this approaches $\rho^2 \geq 1 - \alpha_*$.

The theorem is a capacity ledger. If one fixed decoder works uniformly on at most $d_\varepsilon^{\max}$ microstates, then covering all microstates in Φ_{micro}, which has size $2^{S_{\mathrm{BH}}}$, requires at least $2^{S_{\mathrm{BH}}}/d_\varepsilon^{\max}$ different decoders. Picking one decoder out of that family costs at least $\log_2(2^{S_{\mathrm{BH}}}/d_\varepsilon^{\max}) = S_{\mathrm{BH}} - \log_2 d_\varepsilon^{\max}$ control bits. The covariant entropy bound says a subhorizon patch with area fraction ρ^2 can store at most about $\rho^2 S_{\mathrm{BH}}$ bits. So the patch can hold the decoder choice only if ρ^2 is large enough.

Proof. Apply Lemma 2 with $\mathcal{M} = \Phi_{\mathrm{micro}}$ so that $|\mathcal{M}| = 2^{S_{\mathrm{BH}}}$. For each operational decoder $D \in \mathcal{D}$, take $\mathcal{M}^{(D)}$ to be a uniform success set at accuracy ε. By definition of $d_\varepsilon^{\max}$ I have $|\mathcal{M}^{(D)}| \leq d_\varepsilon^{\max}$ for every D. The single-observer requirement means the protocol must succeed for every microstate, so the success sets cover Φ_{micro}. Lemma 2 then gives the key budget lower bound

$$\mathsf{I}(v_K) \geq S_{\mathrm{BH}} - \log_2 d_\varepsilon^{\max}.$$

Lemma 1 gives the patch ceiling $\mathsf{I}_{\max}(\mathcal{R}) \leq \rho^2 S_{\mathrm{BH}}$. Since the key must be physically instantiated inside $\mathcal{R}$, I need $\mathsf{I}_{\max}(\mathcal{R}) \geq \mathsf{I}(v_K)$. Combine the inequalities and divide by S_{BH} to get

$$\rho^2 \geq 1 - \frac{\log_2 d_\varepsilon^{\max}}{S_{\mathrm{BH}}}.$$

Finally, insert Proposition 1 to obtain the specialised form with α_* and c_ε. □

Corollary 1 (It takes horizon-scale area to verify a horizon)**.** *Fix any $\rho \in (0,1)$ and fix $\varepsilon \in (0,1)$. Under the assumptions of Theorem 1, if*

$$\alpha_* < 1 - \rho^2$$

then no single patch verifier confined to $A(\mathcal{R}) \leq \rho^2 A_0$ can succeed on all microstates for sufficiently large S_{BH}. Equivalently, if α_* is bounded away from one, any successful verifier must control a patch whose boundary area is an order-one fraction of the horizon area.

Proof. Theorem 1 gives the necessary condition $\rho^2 \geq 1 - \alpha_* - c_\varepsilon/S_{\mathrm{BH}}$. If $\alpha_* < 1 - \rho^2$, then for sufficiently large S_{BH} this condition fails. □

As a concrete example, take $\alpha_* = 0.2$ and neglect the subleading term $c_\varepsilon/S_{\mathrm{BH}}$. A single fixed decoder then works uniformly on at most $2^{0.2 S_{\mathrm{BH}}}$ microstates, that is a code subspace with 20% of the black hole entropy in bits.

If you prefer a concrete count, take $S_{\mathrm{BH}} = 10$ bits. Then there are $2^{10} = 1024$ microstates and one decoder can cover at most $2^{0.2\times 10} = 4$ of them. Covering all microstates therefore needs at least $1024/4 = 256$ decoders and selecting among them costs $\log_2 256 = 8$ control bits.

In general, selecting the correct decoder costs about $0.8 S_{\mathrm{BH}}$ control bits. Theorem 1 gives the requirement $\rho^2 \gtrsim 0.8$ and hence $\rho \gtrsim \sqrt{0.8} \approx 0.89$. In other words, a successful verifier must control a patch whose boundary area is about 80% of the horizon area. Conversely, a patch that accesses only a quarter of the horizon area, meaning $\rho^2 = 1/4$, could only work if $\alpha_* \gtrsim 0.75$.

In black hole plus bath models that realise the Page curve using islands, the two surfaces with areas A_1 and A_2 are the competing quantum extremal surfaces that appear in the generalized entropy

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

minimisation problem. For a given geometry and time slice, A_1 and A_2 can be computed semiclassically, so $\alpha_* = (A_2 - A_1)/A_0$ can in principle be extracted as a function of time. Penington and Almheiri et al compute these competing surfaces in explicit models as part of entanglement wedge reconstruction and the Hawking radiation entropy calculation [9,10]. In this paper I treat α_* as an input parameter because its numerical value is model-dependent.

4.5. Summary Table and Tradeoff Curve

The table below collects the variables.

Quantity	Meaning
S_{BH}	horizon capacity in bits, $S_{\mathrm{BH}} = \frac{A_0}{4\ell_p^2 \ln 2}$
$d_\varepsilon^{\max}$	maximal code subspace dimension per fixed decoder at error ε
α_*	reconstruction fraction, $\alpha_* = (A_2 - A_1)/A_0$
c_ε	error term that depends on ε and does not scale with S_{BH}
$\lvert\mathcal{D}\rvert$	decoder family size, $\lvert\mathcal{D}\rvert \geq 2^{S_{\mathrm{BH}}}/d_\varepsilon^{\max} \geq 2^{(1-\alpha_*)S_{\mathrm{BH}} - c_\varepsilon}$
$\mathsf{I}(v_K)$	key bits, $\mathsf{I}(v_K) \geq \log_2 \lvert\mathcal{D}\rvert \geq S_{\mathrm{BH}} - \log_2 d_\varepsilon^{\max} \geq (1-\alpha_*)S_{\mathrm{BH}} - c_\varepsilon$
ρ	patch size fraction, $A(\mathcal{R}) \leq \rho^2 A_0$
No-go condition	$\rho^2 \geq 1 - \frac{\log_2 d_\varepsilon^{\max}}{S_{\mathrm{BH}}} \geq 1 - \alpha_* - \frac{c_\varepsilon}{S_{\mathrm{BH}}}$

Figure 1 plots the same inequality as a tradeoff curve between α_* and the required patch fraction ρ.

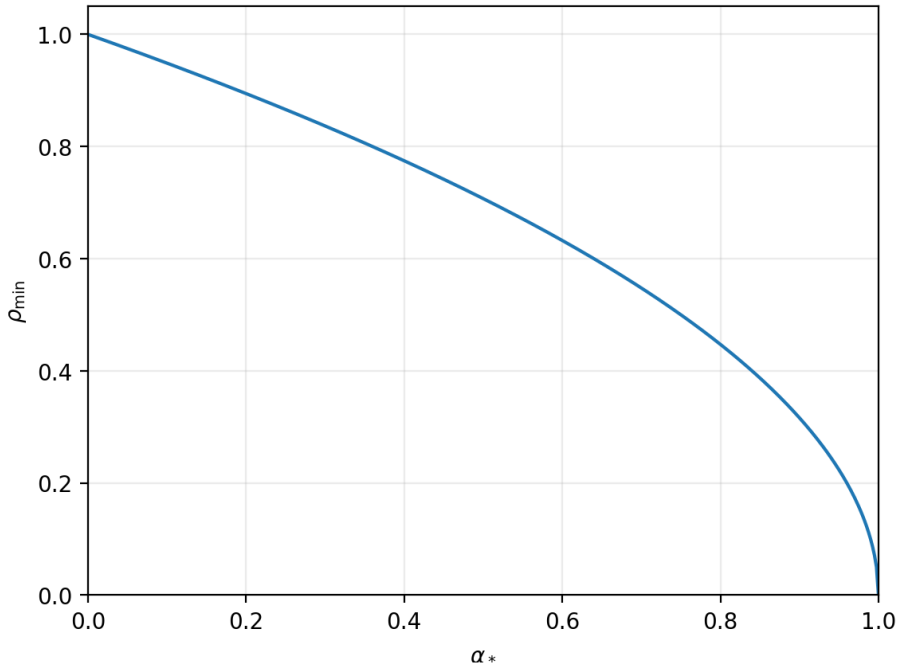


Figure 1. The large entropy limit of Theorem 1 gives a simple tradeoff curve between the reconstruction fraction α_* and the required patch size fraction ρ. In the regime $d_\varepsilon^{\max} \approx 2^{\alpha_* S_{\mathrm{BH}}}$ with subleading constants neglected, the theorem reduces to $\rho^2 \gtrsim 1 - \alpha_*$, so $\rho_{\min} = \sqrt{1-\alpha_*}$. The curve is evaluated on 400 evenly spaced values of α_*. The horizontal axis is α_*. The vertical axis is $\rho_{\min}$. Source data are provided in `source_data/source_data_Fig1_rho_min_vs_alpha.csv`. When α_* is not close to one, $\rho_{\min}$ is an order-one fraction of the horizon, meaning a small infaller patch cannot hold the decoder choice.

5. What the Theorem Does and Does Not Show

Theorem 1 does not prove that firewalls exist. It also does not prove that firewalls do not exist. It proves something narrower and more operational. Under controlled reconstruction assumptions, the

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

standard firewall verification protocol does not fit inside any subhorizon causal patch unless $\log_2 d_\varepsilon^{\max}$ is close to S_{BH}, which corresponds to reconstruction that is close to state independent.

There are of course some potential escape hatches if we violate or make very specific assumptions.

1. Exponential state independence. If α_* is extremely close to one, then one decoder can succeed uniformly on code subspaces whose size in bits is close to the full black hole entropy S_{BH}. That is a strong form of state independence.
2. Horizon-scale control. If α_* is bounded away from one, then Theorem 1 forces ρ^2 to be bounded below by a fixed constant that does not shrink as S_{BH} grows. The verifier must control a patch whose boundary area is an order-one fraction of the horizon area. That is not a small infaller experiment (hence the word "small" in the title of this paper).
3. Give up microstate coverage. If one only demands success on an exponentially small fraction of microcases, the key size can shrink by an amount that scales with S_{BH}. Succeeding on any fixed nonzero fraction still costs order S_{BH} control bits, up to an additive constant. This is not the AMPS claim.

This is not the Harlow–Hayden complexity barrier [12]. Even an oracle for the computation does not remove the need to physically specify which channel is being applied. The obstruction is representational and covariant. It is also not removed by stretching the protocol over more time. In Stack Theory terms, that replaces a demand for co-instantiation at one instant by a weaker demand that each ingredient shows up somewhere in a time window [6,11]. The AMPS paradox is a single-observer contradiction. It only bites if the observer can hold one joint record in which the ingredients are simultaneously present. If the patch ceiling forbids that co-instantiation, then no amount of time sharing can restore the missing joint record.

More broadly, whenever the correct operation depends on hidden microscopic details, the program register can become the scarce resource before the data register does.

6. Conclusion

All I have really done in this paper is translate established generalisation-optimal learning results from Stack Theory into physical language and apply them to the black hole firewall paradox [22], implementing what was already alluded to but not formally established in a conference paper [16].

I have shown that a single observer confined to a subhorizon causal patch faces a capacity bottleneck when attempting the AMPS firewall verification protocol. The core inequality $\rho^2 \geq 1 - \alpha_* - c_\varepsilon / S_{\mathrm{BH}}$ says that if the geometric reconstruction fraction α_* is bounded away from one, then the observer must control a patch whose boundary area is an order-one fraction of the horizon area. The bottleneck is not computational complexity but representational capacity: the physical control register that selects the correct decoder must fit inside the same patch as the verification outcome.

The result is conditional on the assumptions stated in Box 3. The covariant entropy bound is treated as a memory ceiling, which is an interpretation step beyond the formal bound on light-sheet entropy. The state-dependence input relies on holographic reconstruction bounds from evaporating black hole plus bath models. If either assumption is weakened, the quantitative bound changes, but the qualitative point survives whenever one fixed decoder cannot cover all microstates.

Three directions for future work look both feasible and falsifiable. First, island-style reconstruction thresholds can be recast as explicit join constraints, expressing the usual entanglement wedge story as an inequality over capacities rather than a narrative about observers. Second, the universal success requirement in the theorem can be relaxed to a tunable success fraction and matched to known code subspace constructions, turning the key bound into a family of tradeoff curves that can be compared directly to explicit state-dependent reconstruction schemes. Third, the same separation between computation and control should apply beyond black holes. If the covariant entropy bound is a genuine memory ceiling, then similar capacity ledger constraints should appear wherever the correct operation depends on hidden microscopic details, including cosmological horizons and any scenario where the bound is tight.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

Finally, none of these steps require Stack Theory to be a theory of everything. It is merely the road I took to get to this conclusion. Whatever quantum gravity is and whatever formalism we use, quantum gravity should respect this covariant information accounting.

7. Methods and Supplementary Information

GitHub URL: https://github.com/ViscousLemming/Technical-Appendices/tree/main/Papers

All results in the main text are analytic. All the experiments, code, this paper's Supplementary Information (SI) and the supporting Stack Theory community of work is available in the online GitHub repository [6]. Figure 1 is a plot of a closed-form inequality from Theorem 1. Extended Data Figures 1 and 2 are toy computations that illustrate the finite capacity bookkeeping behind Box 2 and Lemma 2. They do not model quantum gravity, holographic dynamics, or black hole evaporation. They are included only for intuition and can be omitted without changing any claim in the main text.

7.1. Supplementary Code 1 in Plain English

Supplementary Code 1 is the Python file `stack_physics_suite_experiments.py`. It produces the PNG figures under `figures` and the CSV source data files under `source_data`. For a plain English description of each PNG and each CSV file, see `figures/README.md` and `source_data/README.md`. The script fixes its random seeds and it strips variable PNG metadata. If you rerun it, you should reproduce the numerical CSV outputs exactly. The PNG files should match on the same plotting stack, but different Matplotlib versions can produce minor byte-level differences.

7.2. How to Reproduce the Toy Computations

Unzip the bundle and run `python -B stack_physics_suite_experiments.py` from the bundle root. The script rewrites three PNG figures and three CSV source data files. The PNG files live under `figures` and the CSV files live under `source_data`. The parameter values used for each plot are stored in the corresponding CSV file and are summarised here.

Output	What it is	Key parameters
Figure 1	theorem curve $\rho_{\min} = \sqrt{1-\alpha_*}$	400 grid points for $\alpha_* \in [0,1]$
Extended Data Fig. 1	toy complementarity histogram	env_n $= 12$, $m = 6$, $p = 0.2$, samples $= 200$, seed $= 2$
Extended Data Fig. 2	toy decoder cover scaling	$\alpha = 0.6$, $n \in \{6, 8, 10, 12\}$, factor $= 12$, trials $= 2$, seed $= 2$

What the toy parameters mean. The parameter `env_n` is the number of bits used to index the toy world states. So the environment has $2^{\text{env_n}}$ mutually exclusive states. The parameter `m` is the number of yes or no questions in each vocabulary. The parameter `p` is the probability that a given world state is included in a given random question. The parameter `n` is the number of microstate bits in the decoder coverage toy model. The toy microstate set has size $|H| = 2^n$. The parameter `alpha` sets the per-decoder coverage size $d_{\max} = \lfloor 2^{\alpha n} \rfloor$. The parameter `factor` controls how many random candidate decoders I generate before greedy cover. The parameter `trials` is how many independent random covers I run at each n. The parameter `seed` is the fixed random seed used for reproducibility.

7.3. How to Read the Plots

This bundle includes one analytic curve and two toy plots. All three are about counting distinctions, meaning how many different internal cases a verifier must be able to tell apart.

Figure 1.

This is the theorem in picture form. Pick a value of α_* on the horizontal axis. That value says how large the best-case decoder code subspace can be, as a fraction of the full microstate entropy. The

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

vertical axis tells you the smallest patch fraction ρ that can hold the decoder choice. If α_* is far below one, then ρ is close to one, meaning the verifier needs access to most of the horizon area.

Extended Data Fig. 1.

This histogram shows a join overflow. Each vocabulary by itself fits under the ceiling. But their join often exceeds the same ceiling. That is the toy analogue of complementarity as a failure of joint realisability under a capacity bound.

Extended Data Fig. 2.

This plot shows why state dependence forces key bits. If one decoder only covers a fraction $2^{\alpha n}$ of 2^n microstates, then you need about $2^{(1-\alpha)n}$ decoders to cover everything. Taking $\log_2$ turns that into a line of slope $1-\alpha$. The greedy set cover curve sits above the lower bound because random decoders overlap.

7.4. Figure 1 in the Main Text

This figure is not a simulation. It plots the large entropy limit of the theorem. In the simplified regime where $\log_2 d_\varepsilon^{\max} \approx \alpha_* S_{\mathrm{BH}}$ and subleading constants are neglected, the theorem gives

$$\rho_{\min} = \sqrt{1-\alpha_*}.$$

I sample 400 evenly spaced values of α_* between 0 and 1 and plot the resulting curve.
Algorithm translated to English.

1. Sample α_* uniformly on a grid between 0 and 1.
2. Compute $\rho_{\min} = \sqrt{1-\alpha_*}$ at each grid point and plot the curve.

7.5. Extended Data Figure 1

The point of this plot is to show complementarity as a capacity overflow in a finite toy model. I build a toy world with $|\Phi| = 2^{12}$ mutually exclusive states. I sample two vocabularies v_1 and v_2. Each vocabulary contains $m = 6$ random programs. Each program is a random subset of Φ obtained by including each state independently with probability $p = 0.2$. For each sampled pair, I form the join vocabulary $v_1 \vee v_2$ by taking the union of their programs. I compute $\mathsf{I}(v_1 \vee v_2)$ and plot a histogram over 200 samples. The illustrative ceiling in the figure is 9.8 bits. The random seed is fixed to 2.
Algorithm translated to English. A histogram is a bar chart of counts grouped into bins.

1. Fix the toy parameters env_n, m, p, the number of samples, and the random seed.
2. For each sample, draw two random vocabularies v_1 and v_2 on the same finite environment Φ.
3. Form the join vocabulary $v_1 \vee v_2$ by taking the union of their programs.
4. Compute the join encoding capacity $\mathsf{I}(v_1 \vee v_2) = \log_2 \left|\mathrm{Im}(\mathrm{Enc}_{v_1 \vee v_2})\right|$.
5. Plot the histogram of these values and draw the ceiling line.

The horizontal axis in the histogram is the join capacity $\mathsf{I}(v_1 \vee v_2)$ in bits. The vertical axis is the number of sampled vocabulary pairs that fall in each bin. The vertical reference line is the illustrative ceiling.

7.6. Extended Data Figure 2

Warning about notation. The toy exponent parameter α used in this section is not the same object as the theorem parameter α_*. The theorem parameter α_* comes from competing quantum extremal surfaces. The toy parameter α only controls the integer subset size $d_{\max} = \lfloor 2^{\alpha n} \rfloor$ in a finite set cover model.

The point of this plot is just to illustrate the decoder counting step in a finite toy model. I model the microstate basis as a finite set H with $|H| = 2^n$. I assume one decoder can succeed on at most $d_{\max} = \lfloor 2^{\alpha n} \rfloor$ microstates. Covering all microstates then requires at least $\lceil |H|/d_{\max} \rceil$ decoders. I plot $\log_2 \lceil |H|/d_{\max} \rceil$ as a function of n.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

I also plot a random overlap model. In that model, each decoder is a random subset of size $d_{\max}$ and I apply a greedy set cover algorithm. Set cover means choosing a collection of decoders whose union contains every microstate in H. Greedy means I repeatedly pick the decoder that covers the largest number of microstates that are still uncovered. Because random subsets overlap, the greedy cover typically needs more decoders than the lower bound. The parameter values in Supplementary Code 1 are $\alpha = 0.6$, $n \in \{6, 8, 10, 12\}$, factor 12, trials 2, and random seed 2. Factor sets how many random decoders I generate in the toy model, relative to the lower bound. Specifically I generate $M = \text{factor} \times \lceil |H|/d_{\max} \rceil$ random subsets before running greedy cover.

Algorithm in plain English.

1. Choose α, a list of microstate bit sizes n, and the random seed.
2. For each n, set $|H| = 2^n$.
3. Set the per-decoder coverage to $d_{\max} = \lfloor 2^{\alpha n} \rfloor$.
4. Compute the no-overlap lower bound $L = \lceil |H|/d_{\max} \rceil$ and record $\log_2 L$.
5. Generate $M = \text{factor} \times L$ random candidate decoders, each as a subset of H of size $d_{\max}$.
6. Run greedy set cover to select a collection whose union is all of H. Record the cover size and average it over the requested number of trials.

The Supplementary Information contains a proof of the capacity complementarity inequality, a formal separation between single patch and distributed realisation following the Temporal Gap non-commutation result, Theorem 3 in [11], and robustness checks for the AMPS ledger. It also documents the toy computations in more detail and explains how each plot connects back to the inequalities used in the main theorem.

Provenance.

The central technical result is Theorem 1 and its corollaries. It packages the decoder selection problem into a simple covariant capacity ledger inequality. The argument combines a covariant area-based capacity ceiling [7] with a state-dependence bound for interior reconstruction [8] and standard limits on programmable quantum operations [13,14]. Related uses of programmable processor bounds together with gravitational entropy bounds to constrain bulk computation in AdS/CFT appear in [15]. The present paper applies this control information lens to firewall verification and combines it with alpha-bit coverage bounds to obtain an explicit horizon area fraction requirement. The Stack Theory primitives used to formalise capacity and control information follow [5,6,16,23]. All plotted data are generated from closed-form expressions or from synthetic toy models in Supplementary Code 1.

Supplementary Materials: The following supporting information can be downloaded at the website of this paper posted on Preprints.org.

Data Availability Statement: No empirical datasets were generated or analysed. Source data for Figure 1 and Extended Data Figures 1 and 2 are provided as CSV files in the bundle under `source_data`. These CSV files are generated by Supplementary Code 1.

Code Availability: As stated earlier the code is in the folder named after this paper in the GitHub repository [6]. Supplementary Code 1 (`stack_physics_suite_experiments.py`) reproduces Figure 1 and Extended Data Figures 1 and 2 and writes the corresponding source data CSV files. It depends on `numpy` and `matplotlib`. A vendored copy of `stacktheory_suite` (version 0.8.4) is included in the top-level directory of the GitHub, which you'll need to re-run the experiment code.

AI-Assisted Technologies: In the preparation of this manuscript I used ChatGPT, Claude, Grok and Gemini for language editing, refactoring code, checking for errors, generating feedback / simulated reviews, and for detailed searches into less familiar topics. The actual substance of this research was from a human, as evidenced by discussion of these ideas in an earlier conference paper on which this paper substantially depends [16]. All I've really done here is translate my existing work into a new context. All theoretical claims, derivations, and final wording were reviewed and approved by me. I take full responsibility for the content.

Author Contributions: I (M.T.B.) conceived the study, developed the theory, and wrote the manuscript.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

Conflicts of Interest: I declare no competing interests.

Correspondence and Requests for Materials: Correspondence and requests for materials should be addressed to M.T.B. (m@michaeltimothybennett.com).

References

1. Susskind, L.; Thorlacius, L.; Uglum, J. The stretched horizon and black hole complementarity. *Physical Review D* **1993**, *48*, 3743–3761, [hep-th/9306069]. https://doi.org/10.1103/physrevd.48.3743.
2. 't Hooft, G. On the quantum structure of a black hole. *Nuclear Physics B* **1985**, *256*, 727–745. https://doi.org/10.1016/0550-3213(85)90418-3.
3. Almheiri, A.; Marolf, D.; Polchinski, J.; Sully, J. Black holes, complementarity, or firewalls. *Journal of High Energy Physics* **2013**, *2013*, 062, [arXiv:hep-th/1207.3123]. https://doi.org/10.1007/jhep02(2013)062.
4. Hausmann, L.; Renner, R. The firewall paradox is Wigner's friend paradox **2025**. [arXiv:quant-ph/2504.03835]. https://doi.org/10.48550/arXiv.2504.03835.
5. Bennett, M.T. How To Build Conscious Machines. PhD thesis, The Australian National University, 2025. https://doi.org/10.25911/ta58-p428.
6. Bennett, M.T. Technical Appendices, 2025. Archived release on Zenodo. Source repository: https://github.com/ViscousLemming/Technical-Appendices, https://doi.org/10.5281/zenodo.17164785.
7. Bousso, R. The holographic principle. *Reviews of Modern Physics* **2002**, *74*, 825–874, [hep-th/0203101]. https://doi.org/10.1103/revmodphys.74.825.
8. Hayden, P.; Penington, G. Learning the Alpha-bits of Black Holes. *Journal of High Energy Physics* **2019**, *2019*, 007, [arXiv:hep-th/1807.06041]. https://doi.org/10.1007/jhep12(2019)007.
9. Penington, G. Entanglement wedge reconstruction and the information paradox. *Journal of High Energy Physics* **2020**, *2020*, 002, [arXiv:hep-th/1905.08255]. https://doi.org/10.1007/jhep09(2020)002.
10. Almheiri, A.; Hartman, T.; Maldacena, J.; Shaghoulian, E.; Tajdini, A. The entropy of Hawking radiation. *Reviews of Modern Physics* **2021**, *93*, 035002, [arXiv:hep-th/2006.06872]. https://doi.org/10.1103/revmodphys.93.035002.
11. Bennett, M.T. A Mind Cannot Be Smeared Across Time. In Proceedings of the Forthcoming in Proceedings of the AAAI 2026 Spring Symposium on Machine Consciousness: Integrating Theory, Technology, and Philosophy, Burlingame, California, USA, April 2026.
12. Harlow, D.; Hayden, P. Quantum computation versus firewalls. *Journal of High Energy Physics* **2013**, *2013*, 085, [arXiv:hep-th/1301.4504]. https://doi.org/10.1007/jhep06(2013)085.
13. Nielsen, M.A.; Chuang, I.L. Programmable quantum gate arrays. *Physical Review Letters* **1997**, *79*, 321–324, [quant-ph/9703032]. https://doi.org/10.1103/physrevlett.79.321.
14. Kubicki, A.M.; Palazuelos, C.; Pérez-García, D. Resource Quantification for the No-Programming Theorem. *Physical Review Letters* **2019**, *122*, 080505, [arXiv:quant-ph/1805.00756]. https://doi.org/10.1103/physrevlett.122.080505.
15. Kubicki, A.M.; May, A.; Pérez-García, D. Constraints on physical computers in holographic spacetimes. *SciPost Phys.* **2024**, *16*, 024, [arXiv:quant-ph/2304.09900]. https://doi.org/10.21468/SciPostPhys.16.1.024.
16. Bennett, M.T. Is Complexity an Illusion? In Proceedings of the 17th International Conference on Artificial General Intelligence. Springer, 2024, Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-65572-2_2.
17. Bekenstein, J.D. Universal upper bound on the entropy-to-energy ratio for bounded systems. *Phys. Rev. D* **1981**, *23*, 287–298. https://doi.org/10.1103/physrevd.23.287.
18. Bennett, M.T. Computational Dualism and Objective Superintelligence. In Proceedings of the 17th International Conference on Artificial General Intelligence. Springer, 2024, Lecture Notes in Computer Science. https://doi.org/10.1007/978-3-031-65572-2_3.
19. Bennett, M.T. Are Biological Systems More Intelligent Than Artificial Intelligence?, 2024, [arXiv:cs.AI/2405.02325]. In press, 2026, Philosophical Transactions of the Royal Society B: Biological Sciences. Special issue on Hybrid agencies: crossing borders between biological and artificial worlds, https://doi.org/10.48550/arxiv.2405.02325.
20. Bekenstein, J.D. Black holes and entropy. *Physical Review D* **1973**, *7*, 2333–2346. https://doi.org/10.1103/physrevd.7.2333.
21. Hawking, S.W. Particle creation by black holes. *Communications in Mathematical Physics* **1975**, *43*, 199–220. https://doi.org/10.1007/bf02345020.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.

Preprints.org (www.preprints.org) | NOT PEER-REVIEWED | Posted: 4 March 2026 doi:10.20944/preprints202603.0227.v1

22. Bennett, M.T. Optimal Policy Is Weakest Policy. In Proceedings of the Artificial General Intelligence. Springer, 2025, Vol. 16057, *Lecture Notes in Computer Science*, pp. 43–56. https://doi.org/10.1007/978-3-032-00686-8_5.
23. Bennett, M.T. The Optimal Choice of Hypothesis Is the Weakest, Not the Shortest. In Proceedings of the 16th International Conference on Artificial General Intelligence. Springer, 2023, Lecture Notes in Computer Science, pp. 42–51. https://doi.org/10.1007/978-3-031-33469-6_5.

Disclaimer/Publisher's Note: The statements, opinions and data contained in all publications are solely those of the individual author(s) and contributor(s) and not of MDPI and/or the editor(s). MDPI and/or the editor(s) disclaim responsibility for any injury to people or property resulting from any ideas, methods, instructions or products referred to in the content.

© 2026 by the author(s). Distributed under a Creative Commons CC BY license.