12 FIREWALL DAYS, 2012–2015

12.1 SCHRODINGER'S CAT IN A BLACK HOLE

If Schrodinger's cat were behind the horizon of an AdS black hole, not yet fallen to the singularity, could we determine its state by a measurement in the dual CFT?[1] A gauge/gravity dualist would naturally answer "yes." The CFT is a complete description of the dual black hole, so this information should be available. Indeed, with my students Heemskerk and Sully, and my colleague Don Marolf, we showed how to get it.[2]

The basic idea, integrating the field equation in the AdS bulk, was worked out by Bena and others noted earlier. It had been refined in a series of papers by Hamilton, Kabat, Lifschytz, and Lowe (HKLL). Our group was the first to apply it to a normal (one-sided) black hole, but it seemed to work just fine. You take the

operators inside the black hole, and integrate first backward through the horizon and then spatially to the boundary to get the CFT operators.

There were various subtleties, the most notable perhaps being that some of our construction required boundary operators out of time order, which mapped to *time folds* in the bulk. We were not thinking about chaos at the time, but now these play a wide role. A second version of the paper had two additions. It was at this point that Sully joined and added the appendix working out the

PROBING THE BLACK HOLE INTERIOR

In the presence of a black hole in AdS, the precursor—the time-evolved boundary local field—is dual to a particle that falls through the event horizon. Hence, measurements of this precursor are sensitive to the physics of the black hole interior.

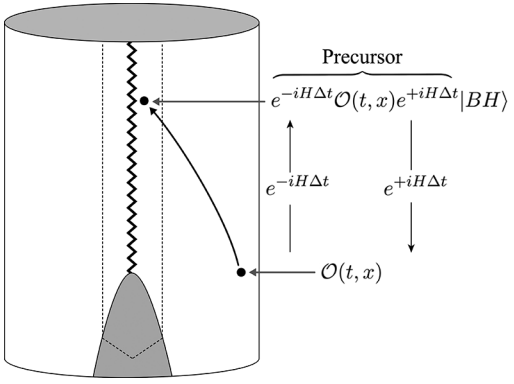

Figure 12.0a

Green's function in detail. And we added a note that the construction worked only before the Page time, because of the firewall, which had just been found.[3]

12.2 BITS, BRANES, BLACK HOLES

In the spring of 2012, KITP ran a ten-week workshop on *Bits, Branes, and Black Holes*. This was directed at the basic questions of quantum gravity: the emergence of spacetime, the connection of area with entropy, the black hole information problem, and so on. For me, I would have said that the central problem was to find the nonperturbative construction of the bulk theory, with the black hole information problem as a key clue.

At the beginning of the program, Ted Jacobson and I were asked to present our perspectives on the information problem. I presented it much as I have noted in the last chapter: gauge/gravity duality showed us that information was not lost, and black hole complementarity (BHC) showed us there was no paradox. But there was still the problem of finding Hawking's mistake: how exactly does the information get out?[4] I thought that what I had said was common lore, but I was surprised. The conservation of information was indeed nearly universally held. But the second part, black hole complementarity, was not. Perhaps the widest response on this was simply not knowing precisely what BHC meant. So I set as my immediate goal to make a simple model of BHC that would answer this.

There were some nice models of the quantum mechanics of black holes that seemed useful here. These *bit models*, due to Samir Mathur and Giddings, were just bits in a line, with rules

for what happens when a bit evaporates from the black hole. These simple models showed the original paradox: either information could not escape from the black hole or it had to travel much faster than light. So I sent my now-seasoned students, Sully[5] and Almheiri, to find a bit model that accounted for BHC. The idea was to break up the bit system into smaller systems, each of which was as much as a single observer could see, and with some kind of junction condition between them. But this failed almost at once.

The problem was that there was a single observer who could see both copies of the information, the one inside and the one outside the black hole, thus violating QM. The original thought experiments that went into BHC had seemed convincing, but a striking paper by Hayden and Preskill, bringing in ideas from quantum information theory, led people to think more clearly about the possible measurements that can be made. So my students and I became more puzzled each week. I was certain that such a basic violation of black hole complementarity must be ruled out, and surely someone at the KITP program could straighten us out. But no one could.

In fact, our own colleague, Don Marolf, had come to the same conclusion by a somewhat different route, thinking about *mining* the black hole rather than just throwing things into it. So Almheiri, Marolf, Polchinski, and Sully (AMPS) joined forces. The fact that we had come to the same result by different arguments, and that no one could easily rebut it, increased our faith in it. Eventually, we tested it on two of the originators of BHC, Preskill at the program conference and Susskind by email. I put off contacting Susskind because I expected the response, "Yes, I thought about this ten years ago, and here is what you're missing"—I had gotten that

from him on other points before. But Preskill and Susskind had the same reaction that we had had: first "this can't be true," and then realizing a week or two later that there was no easy rebuttal.

So we wrote up our results. Three of the principles of BHC cannot all hold: (1) Hawking radiation ends up pure; (2) there was no drama (violation of effective field theory) outside the horizon; and (3) there was no drama for an observer falling through the horizon. So what gives? Not (1): None of us thought that this gave any evidence for information loss, given the problems. My conservative inclination was that some subtle breakdown of effective field theory at distances of the order of the black hole scale would fix things, violating (2). Marolf was sure this did not work, and there had to be drama at the horizon, which he named the *firewall*, violating (3). I tried to make models of how (2) might break down, but I failed. So I had to go along with Marolf's conclusion, that perhaps the most conservative resolution was that the infalling observer *burns up* at the horizon. Another intuition he had was that the interior stopped when the black hole's *quantum memory* became full. So perhaps a *bit wall* would have been more accurate.

FIREWALL

The conditions of unitarity and the emptiness of space at the horizon impose two conflicting conditions on the radiation. The condition of unitarity demands that, at some stage, the emerging Hawking particles be fully entangled with the early Hawking radiation far away in order for the total radiation entanglement entropy to decrease. On the other hand, the emptiness of space at the horizon demands that the outgoing particles be fully entangled with their interior partners.

These statements cannot hold simultaneously because quantum entanglement is monogamous; a particle cannot be fully entangled with two separate systems.

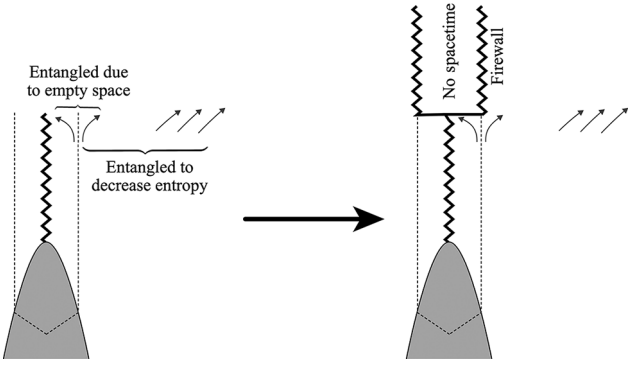

Figure 12.0b

Since the conflict arises when trying to stitch together a simultaneous picture of the inside and outside of a black hole, one could wonder if black hole complementarity could help. Unfortunately, it doesn't: the impossible pattern of entanglement can be detected by a single infalling observer. As crazy as it might seem, the most reasonable way out—at least to Joe and the rest of AMPS—was to break the entanglement across the horizon resulting in a wall of highly excited particles, a *firewall*. This suggests that there may be "no spacetime" behind the event horizon of a black hole.

12.3 PERSONAL NOTES

Before continuing with the physics, a few personal notes.

The three times that I shook up the field—D-branes, the string multiverse, and firewalls—might give you the impression that I am a radical, but it is not by design. Rather, I think I am more like Dirac, with a knack for how theory fits together and the philosophy of "one must be prepared to follow up the consequences." But of course we know that even Dirac did take some time to accept what he had predicted, and so did I. It took me nearly ten years of playing with D-branes before recognizing their importance. And with both the multiverse and the firewall, my inclination was to soft-pedal the results, and it took brilliant young collaborators, Bousso and Marolf, to push things forward.

The second note is a mention of the many others who have proposed modifications of the black hole interior. Chapline, Hohlfeld, Laughlin, and Santiago, and Mazur and Mottola, made such arguments, but I don't think their physics made sense. Braunstein had an argument and conclusions resembling ours, but his black hole Hilbert space was not correct. But the one who certainly did something correct, and important, was Samir Mathur.

Mathur had devoted a large part of his career to the information problem, even after most string theorists accepted *AdS/CFT + BHC* and moved on. He was most known for the idea of *fuzzballs*, modifications of the black hole horizon from higher-dimensional brane configurations. It was proposed that this was the resolution of the information problem. The issue for me was that almost all his arguments were based on nearly supersymmetric black holes. It did not seem that there was any extension to

Schwarzschild black holes. But along the way, Mathur sharpened the information paradox.

In particular, he originated the argument that black hole complementarity violates strong subadditivity of the entropy, which was one of the arguments that AMPS gave. I am sorry that our first and second versions did not acknowledge him on this; I think that because we found his central story about fuzzballs unpersuasive, we did not pay careful enough attention to the rest. Indeed, you might wonder what is the difference between a *fuzzball* and a *firewall*. What we had in mind was the horizon ending as a sea of bits, rather than some geometric structure that extends further.

12.4 FOLLOWING UP

It was fun to have once again kicked over the hive and watched the bees swarm. Though I was a bit peeved that, after we had spent three months looking for flaws, within two weeks people were writing papers explaining why we were wrong without having fully thought through our arguments. But happily the best of them, Raphael Bousso and Daniel Harlow, each recognized their error and withdrew their paper. Susskind did the same, then changed back again, and by now is out on some perpendicular axis.

Of course, we only had an argument by contradiction, not a proof or even a calculation of what happens in the interior. Even on the question of what time the firewall forms, we had only an upper bound, the Page time.[6] I had no good ideas for this, so I spent the next year or more reading what everyone else wrote in response to us. Various alternatives to the firewall were developed

and ideas were exchanged, with a workshop every few months: Stanford, CERN, KITP.

AMPSS, the original AMPS group plus Douglas Stanford, a KITP grad fellow from Stanford, wrote a follow-up in which the arguments were clarified and sharpened. We also pointed out problems with various alternatives to the firewall. One advance was to put the black hole in AdS, where the boundary conditions gave greater control. Black holes normally do not decay in AdS space, but by coupling to an additional heat bath (as in the earlier work of my student Rocha) one could do controlled thought experiments. We also found a simplified argument for the firewall. In its original form a very fast quantum computer was needed. We showed that even without this there was a paradox, using the butterfly effect.

In a second follow-up, Marolf and I added some new arguments and observations. There was a common question: does the firewall invalidate the calculation of Hawking radiation? The usual calculation does depend on the geometry behind the horizon, but causality would seem to say that events behind the horizon could not affect the radiation. By a statistical argument, we showed that the radiation was unaffected. We also returned to the Schrodinger's cat question. We showed that the firewall argument also made it impossible to see what is behind the horizon of an AdS black hole in the dual CFT. This seems to contradict the general assumption about the CFT seeing the whole interior, but makes sense if spacetime ends at the firewall.

What surprised me was how many were willing to modify quantum mechanics in order to avoid the firewall. QM was not one of the explicit assumptions of black hole complementarity; it was

implicit. So I thought of this as a new alternative, *quantum drama*. It differed from Hawking's modification of QM, which was visible in measurements outside the black hole. These, including final state conditions (Lloyd and Preskill),[7] limits on quantum computation (Harlow and Hayden), ER = EPR (Maldacena and Susskind), and state-dependent observables (Papadodimas and Raju),[8] could be restricted to observations behind the horizon. It is possible that one of these is true, but there are issues with each. I particularly had an issue with state-dependence. It sounds benign: aren't observables always state dependent? Well, not in this way, which required that the Born rule of quantum mechanics be modified. So Don and I wrote another paper, making this clear.

Any of these modifications of quantum mechanics might turn out to be correct, but if you are going to modify QM you have a lot of explaining to do. Of course, the other alternative, a modification of the geometry, also needs a lot of explaining. There were various ways this might be implemented: fire (AMPS), fuzz (Mather), strings (Silverstein), *nonviolent nonlocality* (Giddings). As I have noted before, I am a natural agnostic, willing to examine any possibility.

12.5 BRANES

The black hole information problem still seemed like the best insight into the nature of quantum gravity, but after a while the issues seemed to solidify. Perhaps we had to wait for a new insight, as with AdS/CFT. So I was ready for a break. Happily, D-branes still had their puzzles.

My collaborator Karch was still working on AdS/CM, and his student Sichun Sun came to KITP as a grad fellow with a puzzle.

Consider intersecting 0123 and 0145 D3-branes. The 01 intersection degrees of freedom carry a $U(1)$ charged scalar. Duality then requires also a $U(1)$ magnetically charged degree of freedom, but where could it come from: was it an independent field on the intersection, or a solitonic monopole? Neither seemed to make sense: there was no independent magnetic degree of freedom on the branes, but a $1+1$-dimensional intersection did not seem to leave room for a $3+1$-dimensional magnetic soliton.

So, together with my latest student Eric Mintun, we figured this out in four steps: (1) the $N=2$ implied that on a $1+1$-dimensional intersection the scalar couples to the magnetic dipole in addition to the electric potential; (2) the $1+1$ dimensions allowed higher-dimensional interactions, and SUSY required them, thus leading to a $1+1$-dimensional soliton; (3) there was a log divergence in the classical action, which could be treated by the usual process of renormalization (always nice to learn new wrinkles in renormalization, this from Goldberger and Wise); and (4) because of the log, the effective field theory did not make sense up to infinite energy—one needed the branes. So lots of cool field theory in a simple system.

Thinking about this new application of renormalization led to clarifying an old puzzle of mine. The motion of a brane depends on its interactions with other branes. But how does one treat the self-interaction, which is often divergent? What is often done is ignore it, introducing the notion of a probe brane. This was not a controlled approximation. But having understood the classical renormalization of branes, it became clear that they should be understood in the language of effective field theory, with no probe approximation needed. And as I was working this out with

Mintun and an excellent undergrad Philip Saad, the perfect application came along.

Having followed the development of the KKLT model, and participating in part of it, I was puzzled by claims that it was unstable. I tried to understand the arguments (which, incidentally, were largely due to my own former students Bena and Graña) but I could not. So when their student (my grand-student) Andrea Puhm came to Santa Barbara as a postdoc, I tried once again to understand the issues. And they were nicely resolved by the new interpretation of branes: there was no way for a dangerous singularity to arise. So Mintun, Puhm, Saad, a new student Ben Michel, and I wrote up both the correct interpretation of branes and the stability of KKLT.

Getting involved in KKLT led to much more correspondence, and eventually an invitation to speak at SUSY15. Many people had arguments, or intuitions, that these de Sitter vacua could not exist. The stakes were high. If string theory had no such vacua, perhaps string theory was wrong. If it had too few vacua, perhaps the anthropic argument would be ruled out. But looking at the objections, most of them were clearly wrong; some appealed to no-go arguments that were known to be irrelevant even when they were first written down. The most interesting objection had to do with the fact that the KKLT construction required both 10-dimensional and four-dimensional analysis. By careful treatment of scales, we showed that this could be justified in effective field theories.[9]

One more brane puzzle began with Michel studying different duality frames in string moduli spaces. When Puhm joined us, she brought the Saclay point of view on fuzzballs as well as KKLT. In hearing about the simplest (two-charge) case, we realized that the

duality frames had not been fully taken into account. So this became a nice exercise for Michel, Puhm, another postdoc Fang Chen, and me: the "journey to the center of the fuzzball." We followed the different duality frames as we went down the throat, ending up with a geometry different from that previously assumed.[10] Sticking to the highly supersymmetric two-charge geometry meant that we were not close to addressing the fundamental questions, but perhaps we learned something that will be useful down the road.

As an aside, some speakers will refer to their work as a game. I have never liked this: physics is never a game for me. Everything is directed at the big questions, even if circuitously. This is why I am in this field, not to play games. And why is the public paying us?

12.6 PRECURSORS AND CHAOS

Though the firewall puzzle was largely on hold for me, there were many related questions to follow now. Some were motivated by the firewall, but many were motivated by the AdS/CM connection, and by ideas from quantum information. The discovery of the Ryu-Takayanagi (RT) formula generated an enormous wave of interest in the relation between entropy and geometry. Ryu and Takanagi did their work at KITP, and even came into my office at an early stage to ask what it might mean. But having little intuition for entropy, and perhaps some skepticism about the result, I was of little help, and I missed my chance to be an early adopter. Actually, I have not worked on RT yet. Many people jumped into it, and I avoid doing things that other people could do. Perhaps I will wait until they move on, and then look around for what might have been missed.

So my last paper with Almheiri was motivated by AdS/CM, but ended up having some relevance to quantum gravity—it's all connected. I had been puzzled by $0+1$ conformal theories, which often came up in finite density systems. When the transverse directions were compact, the symmetry implied that the low energy density of states had to be of the form $A\delta(E) + B/E$. But the A term comes only from zero energy, and the B term has a divergence and can't continue down to zero energy. So how could there be dynamical states in such systems, as there seemed to be? So we looked at a simple model, based on the CGHS model.[11] I did not work on the first wave of CGHS, twenty years earlier, so I was happy to get a chance to study it. What we found was that the interactions broke the conformal symmetry. This has some current relevance because it happens in the SYK model.[12]

The bulk to boundary operator map was an ongoing interest for me, most recently in the Schrodinger cat question. I had played with it many times, as had others, but I had a sense that all we were doing was to rewrite the AdS/CFT dictionary of Gubser, Klebanov, Polyakov, and Witten. So I was excited by a paper of Almheiri, Dong, and Harlow, which presented something new, perhaps for the first time in twenty years, casting it in terms of quantum information rather than differential equations. In studying the paper, postdoc Vladimir Rosenhaus, Mintun, and I realized that in their nice toy model the quantum information argument could be rewritten in terms of gauge symmetry. I think now that our result was just a special case, but sometimes one just has to throw one's hat into the ring.[13]

Following up on the firewall paradox, Shenker and Stanford began studying the growth of small perturbations to black holes and the butterfly effect (chaos). This was interesting, and I followed

ADS/CFT AND QUANTUM INFORMATION

The AdS/CFT dictionary organizes information in the CFT in terms of different radial distances from the AdS boundary. Information near the boundary is easily corrupted by simple CFT fields, such as local fields, while information deep in the bulk can only be corrupted by complicated precursors. This behavior is reminiscent of quantum error correction, the theory of robustly encoding quantum information to protect it from errors. The emergent holographic radial direction is given a new interpretation as the *level of protection* of the boundary data.

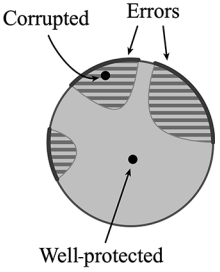

Figure 12.0c

their work for a while until I had an idea of my own. Their papers, like many others, focused on equilibrium black holes, shown by Israel and Maldacena to correspond to two-sided black holes. I was used to the original information problem, with its one-sided state, and so I wanted to see how chaos would manifest there. I quickly realized that it explained something that I had wondered about for a long time.

For more than twenty years, 't Hooft had been presenting what he said was the black hole S-matrix. This did not make sense to

BLACK HOLES, CHAOS, AND THE BUTTERFLY EFFECT

The butterfly effect is when a small change in the past has large consequences in the future. Black holes exhibit this effect in how an outgoing particle just outside the horizon would fail to escape if another particle had been thrown in sufficiently far in the past. The small change of throwing in the particle causes the event horizon to grow and capture the outgoing particle.

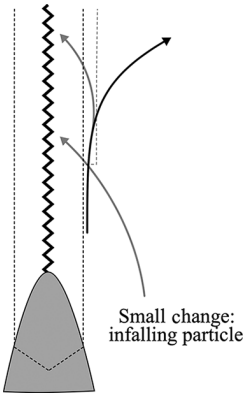

Figure 12.0d

The chaotic nature of a system is measured by analyzing the probability of a large future change as a function of when it was perturbed in the past. The onset of chaos is characterized by an initial exponential growth of this probability, which is controlled by a *Lyapunov exponent*. Black holes have been shown to have the largest possible Lyapunov exponent and are thus the most chaotic physical systems in nature.

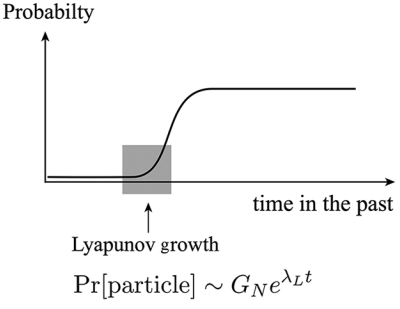

Figure 12.0e

me: it was in the framework of the quantum field theory of GR, with no information from string theory or other completion. But Susskind had told me that one should always pay attention to 't Hooft, so I kept this in mind. In fact what he had calculated was not the S-matrix but the butterfly effect, the change in observables under a small change of state. This could be seen from the time-ordering of the operators. As a side effect, it gave a new and more physical derivation of the firewall.[14]

The subject became more interesting when Alexei Kitaev showed that the chaos in black holes had a characteristic Lyapunov exponent, and that there was a $0+1$ matrix model, the SYK model, that exhibited this. This had some similarity with my old $0+1$ models with Iizuka and Okuda, so with postdoc Rosenhaus, student Michel, and visiting KITP grad fellow Josephine Suh, we looked at whether these models, designed to capture some of the behavior of black holes, might exhibit chaos with the right Lyapunov exponent. Not surprisingly, they did not, being too simple.

Kitaev is famous for not publishing his work, or delaying for years. There was a lot of interest in it, which had appeared only in talks. But Rosenhaus was a dogged calculator, and began to reproduce Kitaev's results, pulling me in. So we obtained the spectrum and four-point functions of the SYK model, reproducing Kitaev's work and getting some new results. Rosenhaus went on with Gross to develop a variety of extensions and variations. They are well matched, liking to talk and calculate for hours on end.

With a new student, Alex Streicher, I was trying to understand the latest from Papadodimas and Raju. This led to a study of the analytically continued partition function. On a trip to Stanford, we found that Shenker and his students were working on the same thing. Eventually, with the addition of numerical types, the group

grew to nine: Cotler, Gur-Ari, Hanada, Polchinski, Saad, Shenker, Stanford, Streicher, and Tezuka.[15]

12.7 WELL, THAT SUCKS

On November 30, 2015, I gave a talk "General Relativity and Strings" at the meeting to celebrate the one-hundredth anniversary of GR. It was held at Harnack House in Berlin, where Einstein often worked and spoke. I was scheduled to speak also the following week in Munich, at a rather different meeting. This was to address whether such theories as strings and inflation were in fact theories. I was looking forward to it; I felt that there were important points that were long overdue to be put forward. My paper, "String Theory to the Rescue," presented the case that string theory, though often criticized, was in fact a great success.

Unfortunately, I never gave the second talk, because three days after my talk at Harnack House I suffered a seizure that sent me to the hospital.[16] I was found to have brain cancer. After many months of surgery, treatment, and recovery, I can write, as you see, but I still do not know whether I will be able to do physics again.

So Rosenhaus finished our last two papers, doing an outstanding job. The other members of the group of nine finished their work and graciously kept my name on the paper, though I was only involved early on. My student Michel developed collaborations with co-advisor Srednicki as well as several other faculty, students, and postdocs, and will be moving to UCLA as a postdoc. My youngest students, Streicher and Milind Shyani, both found new advisors at Stanford, who graciously stepped in. I have always thought highly of Stanford, forming sort of a West Coast axis with us in our interest in the important questions.

Notes

1. I am talking about a one-sided black hole that formed from collapse of ordinary matter, so we know its initial state.

2. I first met Marolf when he was a sixteen-year-old student looking for a grad school in physics. I was happy he chose Austin, though I missed the chance to take him as a student. But it worked out quite well, as he became my colleague at UCSB, in time to do some great work together. But I still think of him as sixteen.

3. Heemskerk had a nice follow-up on this, extending it from scalar fields to gauge fields. When he first took QFT with me, he had a clear interest in the fundamentals, and he got to do some nice work in this area. But he also wanted to make contact with experiments, and I could not promise him that. So after his PhD he moved to biophysics, studying the development of cells with Shraiman.

4. Just to be clear, I like to refer to *Hawking's mistake*, but it is meant to be ironic. He may have been wrong about the answer, but he was right about the importance, and the subtlety, of the question. And his *mistake* has challenged countless theorists for forty years.

5. Sully did not always seem enthusiastic about some of our earlier projects but jumped into emergent spacetime and quantum information. He has done some excellent work since going as a postdoc to Stanford and now to McGill.

6. [The Page time is when the entanglement entropy of the Hawking radiation must start to decrease to satisfy unitarity.—Ed.]

7. [This refers to an elaboration of a proposal by Horowitz and Maldacena that the future singularity of a black hole implements a projection onto the quantum state of the black hole interior.—Ed.]

8. [The idea of state-dependent observables was also proposed during the same period by Erik and Herman Velinde.—Ed.]

9. It is worth noting, though, that without a nonperturbative construction of string theory (aka quantum gravity), the KKLT construction is still a conjecture. It is another argument, beyond the information problem, that we are missing a nonperturbative construction of gravity.

10. We later learned that Martinec and Sahakian had done this first for the zero-spin case, not in the context of the fuzzball.

11. [In fact, this was a rediscovery of a model originally proposed by Jackiw and Teitelboim.—Ed.]

12. Almheiri went to Stanford as a postdoc, and did some remarkable work as mentioned below.

13. Mintun is now a postdoc at British Columbia.

14. I was pleased to learn recently from Stanford that his work with Shenker had been spurred by our discussion of the butterfly effect and the firewall during our work on AMPSS.

15. [This groundbreaking work led to breakthroughs in understanding better what it takes to resolve Maldacena's version of the information paradox. Of the many surprises is how the late-time behavior of the two-point function is governed by the emergence of something akin to Joe's D-branes.—Ed.]

16. David Gross graciously presented my talk, but it was not the same. You can read the original at https://arxiv.org/pdf/1512.02477.pdf, with follow-up at https://arxiv.org/pdf/1601.06145.