Università degli Studi di Urbino Carlo Bo
Department of Pure and Applied Sciences

Ph.D. programme: Research Methods in Science and Technology
Cycle XXXV

A Philosopher's Take on Black Hole Paradoxes

Academic Discipline: Logic and Philosophy of Science (MFIL/02)

Coordinator: Prof. Alessandro Bogliolo
Supervisor: Prof. Gino Tarozzi
Co-Supervisor: Prof. Vincenzo Fano

Ph.D. student: Marco Sanchioni

Academic Year 2021/2022

Contents

Chapter 1

Introduction

This thesis is a philosophical exploration of some of the black hole paradoxes that have guided theoretical physics research in the past two decades.[1] The aim of this thesis is not limited to the mere description of what physicists are doing or a study of the epistemological status of their assumptions and theories, but it goes beyond that. The bigger goal is to actively contribute to the ongoing research on black hole paradoxes by providing a philosophical perspective that can aid in understanding and resolving these paradoxes. I believe that black hole paradoxes are a challenge for physics and have significant implications for our understanding of the nature of space, time, and gravity. Therefore, it is essential to consider the philosophical implications of these paradoxes and how they relate to our understanding of the universe. By examining the philosophical foundations of the theories and assumptions that underlie black hole research, I aim to provide insights that can aid physicists in their ongoing efforts to resolve these paradoxes. Additionally, the study presented in this thesis also seeks to analyse the consequences of the modern theories proposed for the ontology of our world, especially black holes. It is important to note that the goal of this thesis is ambitious, and it may not be easy to achieve. However, I believe that by bringing together the insights of philosophy and physics, we can make progress in understanding these complex and challenging paradoxes.

Before delving into the core of my thesis and providing a philosophical discussion of black hole paradoxes, I would like to take a step back and provide some motivations for my work. First, in §1.1, I will provide my take on philosophy's role in contemporary

[1] In particular, the thesis will be about *firewall* and consequent black hole paradoxes, initially introduced in 2012. Generally speaking, however, black hole paradoxes have been very influential for theoretical physics research since the pivotal Hawking paper (which was published in 1973)

research. In particular, I think that philosophy should be able to interact with open problems by providing a conceptual analysis of them, understanding all the assumptions that are taken into consideration, and showing how specific problems are connected to bigger pictures and are thus connected to other (may be very different) problems. Then, in §1.2, I will tackle the epistemological status of the theories I am considering in my thesis. Indeed, a possible objection to all my work is that I am analysing the conceptual foundation of theories which are mere speculations of high-energy physicists. And, to some extent, this is a good objection. Indeed, none of the theories I present in this thesis has been empirically corroborated. Not even the starting point of all black hole paradoxes, i.e. the Hawking radiation, has been ever detected experimentally. Thus, the worry that we are talking about mere speculation is well-placed. Indeed, my work crucially depends on a particular, narrow, and ultimately unproven conjecture regarding a possible future theory of QG, namely the gauge/gravity duality Maldacena (1999c).[2] How can we draw metaphysical and philosophical lessons from such speculative and uncertain physics? First, let me briefly note that one might object that the dependence on a conjecture renders our view empirically falsifiable. In keeping with the methodology of naturalised metaphysics (Ladyman et al., 2007), empirical relevance should be one of the main features of a proper metaphysical position. From this point of view, what seemed to be a bug in my work, should instead be seen as a feature. While I broadly agree with the perspective of naturalised metaphysics, I take that the criticism articulated above is still excessive. While I certainly agree that the theories I am considering are ultimately still conjectures, they are not nearly as far-fetched as might appear at first sight. In particular, various experiments have been proposed in recent years which might test various aspects of holography and, in particular, of the ER=EPR conjecture

[2] I do not want to go into the details of the assumption of the various black hole paradoxes I am considering since it would be too technical at this stage. All black hole paradoxes are derived into the realm of semiclassical gravity, starting from the work of Hawking (1976). Some of the solutions I present, especially the one connected to the AMPSS paradox, have a strong holographic import. However, it has been shown that such solutions can be derived from a sophisticated procedure on the gravitational path integral, i.e. within the realm of what Wallace (2022) calls perturbative quantum gravity (even though they provided concrete examples of an evaporating black hole in 2-dimension only). Because of these last derivations, the physics upon which I'm building my philosophical discussion should be very general. My philosophical work does not have to be conditionalised on string theory being on the right track. However, how much these last derivations are under control and how much of the results I'm providing should also work for competitive research programs, such as Loop Quantum Gravity, is under scrutiny. And indeed, one of my next works would be to analyse the philosophical foundation of the black hole information paradox solution provided by Loop Quantum Gravity and see how many of the features under study here can be shared with their resolution. Besides this, I will comment on the underlying theory and its possible validity/caveats at each step of the thesis.

(Nezami et al., 2021; Brown et al., 2021; Dai et al., 2020). One of these experiments has already been performed, giving positive results (Jafferis et al., 2022). Moreover, recent advances in quantum black holes have shown that various black hole models display the features characteristic of holography and the ER=EPR (Papadodimas and Raju, 2013; Almheiri et al., 2019a; Penington, 2019; Almheiri et al., 2020a), some of which derived directly from the gravitational path integral (Penington et al., 2019; Almheiri et al., 2020b,a). For these reasons, it seems to me that it is reasonable to begin the project of elucidating some of the philosophical and metaphysical consequences that these theories seem to suggest about the world, keeping in mind, of course, the unproven nature of the conjecture. Indeed, this type of metaphysical project should be understood as trying to articulate what QG's (still speculative) physics might ultimately tell us about reality. Something that has become, if not ordinary, at least broadly accepted in the literature on the metaphysical foundations of QG (Matarese, 2019; Vistarini, 2019; Wüthrich, 2019; Le Bihan, 2020; Jaksland, 2020).

Along these lines, in §1.2, I'm going to argue that the research program centred around the black hole paradoxes can be thought to be progressive, both from a theoretical and an experimental point of view.[3] Indeed, after the advent of the holography principle Susskind et al. (1993) and the AdS/CFT correspondence Maldacena (1999a), I think that the research program in string theory has undergone a drastic shift, which turned a (possibly) regressive research program into a progressive one.[4] And black hole paradoxes are one of the fields in which holography has contributed the most, providing (simplified) models in which one can test features of black holes in quantum gravity.

1.1 Philosophy, Physics, and cutting-edge research

The mainstream part of the theoretical physics community is not convinced that philosophy can help physics. The chapter "Against Philosophy" in a book by Steven Weinberg (Weinberg and Wilczek, 1993), a highly respected physicist, presents a compelling argument that philosophy is more of a hindrance than a help to the advancement of physics.

[3] The content of this section of my thesis is a concise summary of an article which is currently under preparation, and it will thus be underdeveloped.

[4] AdS/CFT came out of the second superstring revolution in 1995, which itself was grounded on a wealth of results, including Strominger-Vafa's calculation of black hole microstates (Strominger and Vafa, 1996; De Haro et al., 2020). Thus, it is worth mentioning that it is controversial where the shift happened. We will discuss in §1.2 why we believe that AdS/CFT has been a true revolution for the field.

Weinberg argues that philosophy can often be a limiting factor for physicists and that they must free themselves from its constraints to make progress. He contends that many of the big questions traditionally in the domain of philosophers are now being explored and answered by physicists, who have the tools and methods to make significant discoveries in these areas. This sentiment is echoed by other prominent figures in the field of physics, such as Stephen Hawking and Neil de Grasse Tyson. Hawking famously stated that "philosophy is dead" because the big questions that were once the province of philosophers are now being explored and understood by physicists. Similarly, Tyson has publicly stated that discoveries in fields such as the expanding universe and quantum physics have rendered the community of philosophers essentially obsolete.

> And so what happens is, the 1920s come in, we learn about the expanding universe in the same decade as we learn about quantum physics, each of which falls so far out of what you can deduce from your armchair that the whole community of philosophers that previously had added materially to the thinking of the physical scientists was rendered essentially obsolete, [...]

These statements, along with Weinberg's argument in the chapter "Against Philosophy", highlight the growing disconnection between philosophy and physics and the perception that the latter has surpassed the former in its ability to answer fundamental questions about the world. However, in the last decades, many important physicists started collaborating with philosophers and thought philosophy could help cutting-edge theoretical physics research. For instance, Carlo Rovelli said:

> Contrary to claims about the irrelevance of philosophy for science, philosophy has always had, and still has, far more influence on physics than commonly assumed.

Another prominent physicist that shared her thoughts on the relevance of the philosophy of physics for contemporary research in theoretical physics is Sabine Hossendfelder:

> I wish more philosophers were willing to make it their business to advance science and to communicate across boundaries. Maybe physicists would complain less that philosophy is useless if it wasn't useless.

This is also my point of view: physics needs philosophy. This is one of the reasons I decided to move to the philosophy of physics: the need to comprehend better the phenomena I was studying and the feeling that the problems I was tackling, especially in the study of black holes in quantum gravity, were more conceptual than technical. However, which is the role of the philosophy of physics in contemporary research in theoretical physics? And, if it has to be a role, why the philosophy of physics did not play such a role in the present time?

Philosophy is valuable for physics as it provides unique methods for gaining new perspectives and critical thinking. Philosophers possess tools and skills that physicists may need to improve, such as the ability to analyse concepts, pay attention to ambiguities, express themselves accurately, identify weaknesses in standard arguments, devise new perspectives, and seek alternative explanations. These skills are not typically part of a physicist's training, but they can be instrumental in advancing the field of physics. Conceptual analysis is one of the most important skills that philosophy provides. Philosophers are trained to examine concepts in great detail, breaking them down into their parts and analysing their relationships with one another. Attention to ambiguities is another skill that philosophy provides. Philosophers are trained to be aware of the many different meanings of words and how concepts can be understood. This skill is instrumental in physics, where technical terms and jargon can often lead to confusion and misunderstanding. The ability to detect gaps in standard arguments, devise radically new perspectives, spot conceptual weak points, and seek out alternative conceptual explanations are all essential skills that philosophers bring to the table. These skills allow philosophers to identify areas where current theories may be lacking and to propose new ways of thinking about a problem. These skills are crucial for advancing the field of physics and for making discoveries. As Einstein was saying

> A knowledge of the historic and philosophical background gives that kind of independence from prejudices of his generation from which most scientists are suffering. This independence created by philosophical insight is—in my opinion—the mark of distinction between a mere artisan or specialist and a real seeker after truth.

As our understanding of the physical world progresses, the questions and problems that we are facing are becoming increasingly complex and multi-faceted. The ability to make precise predictions, which has long been considered a hallmark of science, is no longer sufficient for making progress in these areas. To move forward, we must develop new strategies for identifying and addressing the most important questions and problems. Philosophy can provide valuable insight into this process by helping us identify the assumptions that underlie our current understanding and providing a framework for critically evaluating different approaches to solving problems. One of the most important areas where philosophy can contribute to contemporary research could be quantum gravity. This is a particularly challenging area of research, as it requires a synthesis of our understanding of quantum mechanics and general relativity. These two theories have been developed independently and have yet to be fully reconciled. Philosophers can help identify the conceptual problems that need to be addressed and provide guidance on how to approach these problems in a way consistent with our current understanding of physics. In addition to providing a valuable perspective on the theoretical issues facing physics, philosophy can also contribute to the experimental side of research. For example, philosophers can help identify the key experimental tests that need to be performed to distinguish between different theories and can guide how to design experiments capable of testing these theories in a rigorous and unbiased way. Thus, philosophy can play a crucial role in contemporary research by providing novel perspectives and critical thinking. The methods and skills of philosophers, such as conceptual analysis and attention to ambiguity, can help to identify gaps in current arguments, devise new perspectives, and seek out alternative conceptual explanations. In the current climate of theoretical physics, where theoretical predictions have surpassed experimental capabilities, and the search for a theory of quantum gravity poses both technical and conceptual challenges, the philosophy of physics is well-positioned to make a meaningful contribution to research.

Unfortunately, the relationship between philosophers and physicists has not been as productive as it could be in recent years. This is mainly because the two communities still need to establish a strong working relationship. The lack of collaboration between philosophers and physicists has resulted in missed opportunities for both fields, as philosophers have tools and skills that can greatly benefit physics research and vice versa. To truly progress in physics, philosophers and physicists must work together more closely and establish a strong, collaborative relationship. This will require a change in attitude and approach on both sides, but the potential benefits make it well worth the effort. However, if both communities could overcome these prejudices and work together, the benefits for both sides would be enormous. Philosophers could provide valuable insights and critical thinking skills to guide the development of new theories and experimental designs. Physicists could provide valuable data and experimental results that help philosophers better understand the nature of the physical world. Ultimately, a collaboration between philosophers and physicists would lead to a more complete and accurate understanding of the universe and our place in it.

My thesis addresses an essential element that needs to be added to the current discourse between philosophy and physics. Specifically, it aims to bridge the gap between philosophers and cutting-edge research in theoretical physics, particularly in black holes and quantum gravity. One of the main challenges facing philosophers in this field is the need for a sufficient background in the technical aspects of physics. This can make it difficult for them to engage in critical dialogue with physicists and contribute meaningfully to the ongoing research. My thesis aims to address this issue by providing philosophers with a self-contained introduction to black holes in quantum gravity. This includes a detailed examination of the key concepts and theories and an in-depth exploration of the current debates and controversies surrounding the topic. The focus will be on black hole paradoxes, some of the most challenging and intriguing problems facing physicists today. By providing philosophers with a comprehensive understanding of these issues, my thesis hopes to enable them to actively engage with and contribute to ongoing research in this field. In conclusion, my thesis can be thought to continue the attempts to address the gap between philosophy and physics by providing philosophers with the necessary background to critically engage with cutting-edge research in theoretical physics, particularly in black holes and quantum gravity. It aims to empower philosophers to meaningfully contribute to ongoing research and deepen our understanding of the universe.

1.2 Lakatos Rational Reconstruction and Quantum Gravity

An important objective of the philosophy of science is to establish a demarcation criterion that allows for the clear separation of science from pseudoscience. This work is particularly critical in today's context, where the search for a theory of quantum gravity is ongoing. The scantiness of experiments makes it difficult to establish a solid empirical basis for different theories of quantum gravity. Thus, the ability to distinguish between science and pseudoscience becomes even more important in determining which theories are worth pursuing.

These questions are not purely academic but have long interested some of the most brilliant minds in physics. Indeed, Richard Feynman, in Leonard Susskind's recollection in his book (Susskind, 2008, pp. 119-120), used to be troubled exactly by this problem:

> Feynman said that he had never thought about it. What's more, he had come to dislike the subject of quantum gravity. The effects of Quantum Mechanics on gravity, or gravity on Quantum Mechanics, were just too tiny ever to be measured. It wasn't that he thought the subject was intrinsically uninteresting, but without some measurable experimental effects to guide theory, it was hopeless to guess how it really worked. He said that he had thought about it years ago and didn't want to get started thinking about it again. He guessed that it might be five hundred years before quantum gravity would be understood.

The demarcation criterion can be vital in evaluating the theoretical proposals for quantum gravity and determining their scientific validity. It can help in identifying the theories that are based on sound scientific principles and evidence and those that are not. This process is crucial in identifying the theories worth further investigation and those that should be discarded. Furthermore, it can help avoid wasting resources and time pursuing theories based on flawed reasoning or lack of evidence. In the end, it can lead to a more efficient and productive process of scientific inquiry.

This section can be understood as the beginning of an analysis of the current status of string theory and the holographic principle as leading candidates for a theory of quantum gravity by utilising Lakatos' research program methodology (Lakatos, 1976).[5] Similar work has been done in Johansson and Matsubara (2011), where they analyse the epistemological status of string theory from various perspectives. We intend to build on this analysis by considering recent theoretical and experimental advancements. My main take is that the story of string theory has undergone significant changes in recent years, particularly with the discovery of the holographic principle and its subsequent incarnation into a concrete theory of quantum gravity through the AdS/CFT correspondence (Maldacena, 1999b). These developments have led to new perspectives on the string theory research program and its potential as a theory of quantum gravity. Using Lakatos' methodology, we aim to evaluate the scientific validity and progress

[5] Philosophers of science would disagree that there is only one demarcation criterion. Nowadays, few of them believe that there is a single demarcation criterion. Indeed, Lakatos' sophisticated falsificationism can not be understood as a demarcation criterion but rather a criterion for progress.

of the string theory research program in light of these recent developments. We will examine the empirical predictions, the problems faced by the theory and its ability to solve them, and its ability to generate new theoretical and experimental research. We will also analyse the research program's ability to withstand criticism and its capacity to grow and evolve.[6]

To argue that string theory and the holographic principle constitute a progressive research program, we will present evidence of the development of the literature on string theory and its theoretical and experimental advancements. First, we will examine how publication in string theory changed after the advent of the pivotal paper on the AdS/CFT correspondence (Maldacena, 1999b). Then, we will examine the theoretical achievements of this research program. One of the critical features of a progressive research program is its ability to generate new theoretical research, also known as fecundity. We also show that holography has the potential for unifying different areas of physics into a single theoretical framework. Lastly, we will examine the experimental achievements of holography since another critical aspect of a progressive research program is its ability to make novel predictions that can be tested experimentally.

1.2.1 A String Theory Before and After Holography

Maldacena (1999b) is a seminal paper in modern theoretical physics that has had an enormous impact on the field. It is the most highly cited paper in the history of modern high-energy physics, with an impressive number of citations. According to recent data, the paper has received 21311 citations,[7] which includes references in 1741 thesis, 1696 conference papers, 499 reviews, 234 lectures and 53 books.[8] This is a testament to the far-reaching influence and importance of the paper's ideas and contributions to the field. Its impact can be observed in how it has stimulated further research and development in theoretical physics and has been widely adopted in the scientific community.

These data push the idea that Maldacena (1999b) has been a groundbreaking discovery

[6]Note that, in this section, we use the word holography in a loose (and maybe imprecise) sense. The holographic principle would refer to the statement that the maximal number of microscopic degrees of freedom associated with a spacelike region is proportional to the area A of its boundary in Planckian units. On the other hand, the gauge/gravity duality (or the AdS/CFT correspondence) refers to a precise incarnation of the holographic principle in the realm of string theory. While a very precise differentiation can be made between the two, we will speak loosely in this section by using the two terms interchangeably.

[7]Taken from google scholar.

[8]All other datas are taken from Inspires.

for contemporary research in theoretical physics. One possible criticism of this idea is that its widespread popularity and acceptance may have been driven more by sociological factors rather than scientific merit. It is suggested, for instance, in Smolin (2007), that the paper's success may have been due to a trend or fad in the scientific community, which led to its widespread attention. This has led to continued research and development in this area, not necessarily because it was revolutionary or groundbreaking for the field, but due to its popularity and acceptance within the community. This criticism suggests that the paper's impact may have been more due to a sociological phenomenon than its scientific contributions. The paper's ideas and concepts may not have been as groundbreaking as they were claimed, and its success may have been largely due to external factors such as the social dynamics and trends within the scientific community. However, It is important to note that such criticism should be considered carefully and evaluated in the context of the overall impact and contributions of the paper to the field.

Indeed, the AdS/CFT correspondence has significantly impacted not only string theory but also other disciplines.[9] This fact undermines the idea that its success was solely due to a sociological trend. The AdS/CFT correspondence has been applied in fields such as condensed matter physics, quantum field theory, and quantum gravity, among others, which suggests that it has scientific merit and is not just a popular trend in the scientific community. It has been used to study a wide range of phenomena, such as the behaviour of strongly correlated systems and the thermodynamics of black holes. This shows that the AdS/CFT correspondence can provide new insights and understanding in various fields. However, it is possible to argue that sociological factors may have also driven the success of AdS/CFT in these other fields. In each discipline, it is possible that a similar mechanism has been triggered, where the popularity and acceptance of AdS/CFT have led to its widespread application and use. However, in these fields, AdS/CFT is only a small part of the research; therefore, it is not a dominant trend, making the criticism less strong. It is important to note that the AdS/CFT correspondence is not the only tool used in these fields, and its application is considered alongside other tools and methods.

Another trend that contradicts the idea that AdS/CFT's success was solely due to sociological factors is the increase in AdS/CFT-related publications in Physical Review

[9] A statistical analysis of this claim has been provided in Sanchioni (202x).

Letters, which are highly respected non-specialist journals in physics.[10] These journals have a wide readership and are not limited to string theorists or theoretical physicists. Instead, they are read by a wide range of physicists, including experimentalists and cosmologists. The number of articles in these journals discussing AdS/CFT has grown significantly in recent years. This indicates that AdS/CFT is considered important not only by the mainstream of string theory or theoretical physicists but also is becoming increasingly mainstream among all physicists. This increase in the publication in these non-specialist journals suggests that AdS/CFT's ideas and concepts are recognised as having broad implications and applications beyond string theory. This may be a second reason not to consider the trend only from a sociological point of view. The fact that AdS/CFT is being published in these high-impact journals and being read by a wide range of physicists suggests that it is recognised as a valuable and essential concept. Furthermore, it indicates that the ideas and concepts of AdS/CFT are being considered for their scientific merit and not just because of a sociological trend.

1.2.2 Theoretical Progresses

Holography led to the opening of many new areas of research and has been able to be applied to very different areas of research, including quantum field theory, quantum gravity, and condensed matter physics, contributing to the unification of modern physics. According to Lakatos, these two features, fecundity and unification, are typical of theoretically progressive research programs. The ability of holography to generate new research avenues and connect seemingly disparate fields aligns with Lakatos' criteria for a progressive research program. This makes it a valuable and promising area of study. Fecundity, in this context, refers to the ability of a research program to generate new research questions and open up new areas of investigation. Holography has done this, for instance, by providing new insights into the behaviour of black holes and the nature of space-time, leading to the development of new theories such as holographic duality and the AdS/CFT correspondence. Unification, on the other hand, refers to the ability of a research program to connect seemingly disparate fields and bring them under a common theoretical framework. Holography has done this, for instance, by providing a way to study strongly correlated systems in condensed matter physics using insights from quantum gravity and black hole physics. This has led to the unification

[10] A statistical analysis of this claim has been provided in Sanchioni (202x).

of concepts and ideas from different fields, such as the equivalence of a certain class of condensed matter systems and a certain class of black hole geometries. Thus, one can argue that holography is a theoretically progressive research program that has led to many new discoveries and expanded the boundaries of our understanding of the universe. Its ability to generate new research avenues and connect seemingly disparate fields aligns with the criteria for theoretically progressive research programs outlined by Lakatos. Due to the limited space, we will only mention a few of these many new areas of research that have been opened up by holography, giving a flavour of what holography can do for these diverse fields of research.

Holographic Principle and Quantum Chaos[11]

The holographic principle has been used to study chaotic systems in quantum mechanics. One of the key features of chaotic systems is the phenomenon of the *butterfly effect*, where a small change in the initial conditions leads to a large change in the final state. Holography has been used to study this effect in quantum systems. In particular, the holographic principle has been used to show that the butterfly effect in quantum systems is related to the chaotic behaviour of the dual gravitational theory in the bulk. Another application of holography in the study of quantum chaos is the study of thermalisation. Thermalisation is the process by which an isolated quantum system reaches thermal equilibrium.

Holographic Principle and Condensed Matter Systems[12]

The holographic principle has been used to study strongly correlated systems in condensed matter physics. One of the critical features of strongly correlated systems is the phenomenon of emergent phenomena, where new properties and behaviour appear in a system that its parts cannot explain. Holography has been used to study emergent phenomena in condensed matter systems, particularly phase transitions. Phase transitions are when a system changes from one phase to another, such as from a gas to a liquid. Holography has been used to study this process in condensed matter systems and has provided new insights into the behaviour of these systems.

Holographic Principle and Quark-Gluon Plasma[13]

[11] For a review and recent developments on this subject see (Jahnke, 2019).
[12] For a review on this subject see (Hartnoll, 2009).
[13] For a review and recent developments on this subject see (Gürsoy, 2016).

The holographic principle has been used to study the properties of quark-gluon plasma, which is a state of matter that is thought to have existed in the early universe and can be created in high-energy heavy ion collisions. One of the critical features of the quark-gluon plasma is its high energy density and temperature, leading to quarks and gluons' deconfinement. Holography has been used to study the properties of the quark-gluon plasma under these conditions. In particular, the holographic principle has been used to study transport coefficients related to the transport of energy and momentum in these systems.

Holographic Principle, Quantum Information and Neural Network[14]

The holographic principle has been used to study the connections between quantum information and neural networks. One of the key features of quantum information is the phenomenon of entanglement, where quantum systems become correlated in such a way that the properties of one system cannot be described independently of the others. Holography has been used to study entanglement in neural networks. In particular, the holographic principle has been used to show that a theory of gravity can describe the entanglement of quantum systems in a higher dimensional space-time. In these regards, we mention two possible applications of holography. Concerning quantum information, holography helps in the study of quantum error correction. Quantum error correction is a technique used to protect quantum information from errors caused by noise and decoherence. Regarding the connection between quantum information and neural networks, holography helped study quantum machine learning. Quantum machine learning is a field that combines quantum information with machine learning techniques. Holography has been used to study quantum machine learning algorithms and has provided new insights into the behaviour of these systems.

Besides these vast new research areas, holography has been extremely useful for studying the quantum properties of black holes. We will appreciate the utility of holography in these regards throughout this thesis.

1.2.3 Experimental Progresses

The holographic principle program is not only theoretically progressive but also experimentally progressive. This can be seen through various predictions made by the

[14]For a review and recent developments on this subject see (Gan and Shu, 2017).

program and subsequently confirmed by experiments.[15] As examples, we list two predictions that have been experimentally confirmed, namely, one on the bounds of shear viscosity and another on the derivation of transport phenomena in quark-gluon plasmas. Additionally, at the end of this section, we briefly overview the most recent experiment conducted on the ER=EPR conjecture.

Shear Viscosity Bounds

The shear viscosity is a transport coefficient that measures a fluid's resistance to shear deformation. It is an essential property of fluids, particularly in heavy ion collisions where a Quark-Gluon Plasma is created. The shear viscosity of quark-gluon plasma is of great interest in studying the properties of this state of matter. In the case of strongly coupled $N = 4$ supersymmetric Yang-Mills plasma, the shear viscosity has been studied using the holographic principle (Policastro et al., 2001). One of the critical features of strongly coupled systems is that the shear viscosity is typically much smaller than the entropy density. This is known as the KSS bound. Using the holographic principle, it has been shown that the shear viscosity of strongly coupled N = 4 supersymmetric Yang-Mills plasma saturates the KSS bound. This means that the shear viscosity is much smaller than the entropy density, which is a unique property of this state of matter. Experimental confirmations of this bound, which can be considered holography predictions, are Schäfer and Teaney (2009), Cremonini (2011), and Shen et al. (2011).

Gravitational Anomaly and Transport Phenomena

A gravitational anomaly is a phenomenon that arises in the presence of a gravitational field and is characterised by a violation of the conservation of energy and momentum. This anomaly is related to the presence of a chiral anomaly in quantum field theory, which is a violation of the conservation of chiral charge. The study of gravitational anomaly and its connections to transport phenomena has been an active area of research in recent years. One of the key ways in which the gravitational anomaly is related to transport phenomena is through the study of chiral magnetic and chiral vortical effects. The generation of a current characterises these effects in a magnetic or vortical field in the presence of a chiral imbalance. The study of chiral magnetic and chiral vortical effects has been used

[15] Understanding in which sense these are novel predictions in the sense of Lakatos is a topic which will be discussed in Sanchioni (202x).

to gain insights into the behaviour of strongly correlated systems, such as quark-gluon plasma, created in heavy-ion collisions. In particular, it has been used to study the transport properties of the quark-gluon plasma, such as the electrical and vortical conductivity. The study of the gravitational anomaly and its connection to transport phenomena has also been linked to the study of holographic models. These models use the holographic principle to study the properties of strongly correlated systems and make predictions about their behaviour, such as the transport coefficients (Landsteiner et al., 2011). This is an example of phenomena derived from holography and then experimentally tested in Gooth et al. (2017).

Finally, let me spend a few words on a recent experiment, which can be considered the first step towards the test of the ER=EPR conjecture. ER=EPR is the easiest model of the AdS/CFT correspondence, in which the AdS spacetime can be understood using semiclassical gravity only. And, as we will see, ER=EPR will play an essential role in understanding black hole physics and paradoxes. The experimental test was conducted in 2020 by researchers using Google's 54-qubit quantum computer called *Sycamore* (Jafferis et al., 2022). The experiment in question utilises the holographic principle to observe the passage of a qubit through a two-dimensional traversable wormhole, projected by two entangled SYK models, in a one-dimensional AdS space-time. To fully understand the features of the experiment, it is essential to have a deeper understanding of concepts such as quantum teleportation, the SYK model, and traversable wormholes. Quantum teleportation is a process that uses entanglement between two particles, A and B, to transfer the quantum state of a third particle C, which is in contact with A. An observer measures the system composed of A and C. The measurement results are then communicated to a second observer via classical bits, who uses this information to recreate the state of C by manipulating the state of B.
In 2020, researchers at Sycamore combined quantum teleportation with SYK models developed by Sachdev and Ye (1993). The SYK model, in a nutshell, is a model which describes a system composed of entangled qubits. It was later discovered that the holographic dual of an SYK model is a black hole in a one-dimensional AdS space-time. Based on this, physicists proposed a quantum teleportation protocol using two entangled SYK models, which according to the holographic principle, is equivalent to a qubit passing through a traversable wormhole. Building on that, an experimental method has

been proposed to teleport a qubit between two entangled SYK models by exploiting the rotation of all the spins of the particles constituting the SYK models. This process is the holographic dual of a qubit passing through a wormhole made traversable by a shock wave of negative energy, which in the quantum model represents the spin rotation. Negative energy plays a key role in keeping the wormhole open.

The team of researchers, led by experimental physicist Maria Spiropulu, attempted to replicate the teleportation of a qubit through a traversable wormhole using the Sycamore quantum computer. However, it was impossible to use the full SYK model proposed by physicists, as it would have required ten entangled qubits and 210 connections between them. The simulation would have been too complex for the current technology and needed a quantum computer that is not yet available. The team, therefore, used machine learning methods similar to those used in neural networks to simplify the SYK model to 7 qubits with five connections. This simplified version retained the key properties of the holographic dual necessary for the experiment.

One natural question that arises from the laboratory experiment is whether or not the wormhole "created" in the laboratory is real or not. This is an important question because it has implications for our understanding of gravity and the nature of spacetime. A very first discussion about this issue has been provided in Weinstein (2023). The ER = EPR hypothesis suggests that the way gravity behaves can be explained using concepts from quantum information theory. One example of this is the idea that a traversable wormhole can be thought of as a type of quantum circuit. However, just because this analogy exists doesn't necessarily mean that the wormhole picture of gravity is actually true. The best we can say is that the idea that gravity works through "teleportation-by-size" currently provides the best explanation for the available evidence. What we can safely say is that the wormhole/gravitational picture is equivalent to the teleportation/quantum informational picture, as the authors of Jafferis et al. (2022) also say

> The traversable wormhole expressed as a quantum circuit, equivalent to the gravitational picture in the semiclassical limit of an infinite number of qubits.

However, in order to say something about the reality of the wormhole, one should better understand the ontology of dualities in general and of the ER=EPR conjecture in particular. Investigating such a possibility is a very interesting topic of research. It is important to note that the study of wormholes is still a relatively new field, and much research is needed to fully understand the implications of this experiment and the true nature of wormholes.

1.3 Structure of the thesis[16]

Building on the last two sections, I believe that a conceptual analysis of the black hole paradoxes is extremely important for contemporary research, mainly for two reasons. First of all, the novel resolutions of the black hole paradoxes which have been proposed are (mainly) applications of the holographic principle, which, as discussed in §1.2, can be considered a progressive leading research program in quantum gravity. Second, the study of black holes, especially black hole paradoxes, is one of the spots in which philosophy can contribute to physics since there are also purely conceptual problems to be taken into account. For these reasons, I believe that the content of the present thesis could be of interest to philosophers and physicists and can be considered a first step towards a philosophical analysis of black hole paradoxes.

The first chapter of the thesis (§2) will be dedicated to the analysis of the role of paradoxes in modern theoretical physics. The study of paradoxes in science is an important aspect of the advancement of our understanding of physical phenomena. Paradoxes can be seen as puzzles that need to be resolved in order for a theory to be considered consistent and complete. In this chapter, we will explore how paradoxes can be helpful for several reasons, from theory developments to theory interpretations. We will begin by discussing the conceptual analysis of paradoxes in general. We will examine different types of paradoxes that have been encountered in physics, such as those in quantum mechanics and relativity. We will also discuss the historical context of these paradoxes and their impact on the development of scientific theories. We will argue that the conceptual analysis of paradoxes in general, and black hole paradoxes in particular, can be of extreme importance for exploring new resolutions and gaining a better understanding of the known ones. In particular, we will see how paradoxes can be a powerful tool for driving scientific progress by challenging our current understanding and leading to new insights and discoveries. This will set the stage for the rest of the thesis, where we will delve deeper into black hole paradoxes and their resolutions.

[16]The content of chapters 3, 4, and 5 is part of joint work with Enrico Cinti (Cinti and Sanchioni, 2020, 2021a, 202x).

The second chapter (§3) is a self-contained review of recent black hole paradoxes. As ground-setting material, we will review Hawking's black hole information loss paradox and the Page-time paradoxes. Then, we are moving to the two main paradoxes we analyse in this thesis: the AMPS and AMPSS paradox. Throughout the chapter, we are also reviewing some important physical processes and mathematics relevant to describing the paradoxes, such as the black hole evaporation process, the thermodynamics of black holes, and some ingredients of holography, particularly the program of holographic interior reconstruction.

In the third chapter (§4), I introduce an important philosophical tool that I will be using, which I call *causal structure*. The idea is to recast the inconsistencies that appear in the various paradoxes as inconsistencies between two causal structures: the one representing spatiotemporal relativistic relations and the one representing entanglement quantum mechanical relations. This way, we can elucidate the main point of inconsistency within the various paradoxes and propose logically compatible solutions.

The fourth chapter (§5) is an in-depth analysis of the AMPS and AMPSS paradoxes using the tool of causal structure. We will elucidate the role of two implicit assumptions, called *spacetime distinctness* and *semiclassical exactness*, which will play a major role in resolving the paradoxes. Moreover, we will match the logically compatible solutions with the proposed resolutions in the recent physics literature. Finally, we will see that the same strategy that can be used to solve the AMPS paradox also solves the AMPSS paradox. Thus, both paradoxes can be solved by using the same conceptual move.

Finally, in chapter §6, we are considering the implications that the holographic resolution of the black hole paradoxes can have on the ontology of black holes in quantum gravity. This can be regarded as the very first step in this direction, which, I think, can be an essential and fruitful research program. I will analyse the status of the quantum membrane paradigm, i.e. the idea that a membrane of black hole microstates should be posited at the black hole horizon. If the proposed resolutions of the black hole paradoxes are correct, setting such a membrane would be inconsistent. This, however, is just a negative result: no membrane at the horizon can be posited in the fundamental ontology of black holes. It would be essential to propose a positive metaphysics, in which one argues which is the correct ontological structure of black holes in quantum gravity, given a particular solution to the black hole paradoxes. However, this latter task is left for future work.

Chapter 2

The Role of Paradoxes in Physics

The first chapter of this thesis can be understood as motivating the attention that I, with many physicists working on the foundations of Quantum Gravity, have been paying to black hole paradoxes. In particular, the focus will be on the black hole paradoxes' role in developing a possible theory of Quantum Gravity. I will do so by analysing paradoxes' role in recent research in fundamental physics, especially when theories have overcome our empirical possibilities. It is a fact that, right after the advent of Quantum Field Theory, around half of the past century, we could not immediately test our theoretical hypotheses since we lacked the technological tool needed to test them. Indeed, one of the most famous examples of a theoretical prediction we could not immediately test was the famous Higgs Boson, which was found at the LHC only after 40 years of experimental search. Nowadays, the status of the relationship between theoretical predictions and experiments is even more subtle and complicated. Indeed, it seems that we are not just able to test our physical theory *in practice*, i.e. that we do not yet have the technologies to test them, but rather we are not able to test them *in principle*. Indeed, there is a vast literature on the fact that the fundamental research in theoretical physics concerning Quantum Gravity could be empirically incoherent because of the absence of spacetime in the fundamental ontology of the world. Without spacetime, we cannot do experiments, not even *in principle*, because experiments are necessarily located in space and time.[1]

Even setting aside the possible empirical incoherence of Quantum Gravity, the problem with Quantum Gravity is still that it is currently not possible to perform experiments

[1] See Huggett and Wüthrich (2013) for more on this problem. Note, however, that there are proposals explaining that there is actually no problem (Baron and Bihan, 2022)

that can test its predictions. This poses a challenge for making progress in this field of theoretical physics. The idea that I want to defend in this chapter is that black hole paradoxes can provide a way to gain insight into the nature of Quantum Gravity without relying on experimental data. By studying black hole paradoxes, researchers can gain a better understanding of the underlying principles and structure of Quantum Gravity. It is important to note that black hole paradoxes cannot replace experiments. Empirical confirmation is a crucial step in scientific practice and is necessary for making real scientific progress in Quantum Gravity. However, black hole paradoxes can be a guide towards a more fundamental theory without empirical guidance. They can provide insight and motivation for the development of new theories, which can then be tested and confirmed through experiments. Thus, while experimental data is essential for scientific progress, black hole paradoxes can still be a valuable source of insight into the development of Quantum Gravity. They can provide a way to gain an understanding of the nature of this theory without relying solely on experimental data and can be a guide towards a more fundamental theory that can eventually be confirmed through experiments.

To underlie the significance of studying black hole paradoxes, I will start by analysing what we mean with the word *paradox* when we speak of black hole paradoxes. Generally, a paradox can be understood as a counter-intuitive statement, i.e. a situation that goes beyond common sense. However, one can give a more refined definition of what we mean by a paradox. Indeed, there is a vast literature in philosophy with many possible descriptions of paradoxes. In §2.1, I will propose several classifications of paradoxes meant to underlie the main features of the paradoxes we are interested in when we speak of black hole paradoxes. I will then, in §2.2, provide three different examples of paradoxes, which share the main features of black holes paradoxes, taken from the history of modern physics, namely the *train paradox* (§2.2.1), the *elevator paradox* (§2.2.2), and the *EPR paradox* (§2.2.3). In §2.3, I will underlie the different roles these three paradoxes have played in scientific progress. In §2.4, I will argue which role black hole paradoxes can play in the progress of Quantum Gravity.

2.1 Which Paradoxes?

A paradox is often understood as a self-contradictory or counter-intuitive statement. Paradoxes have been used repeatedly in science, mathematics and philosophy, and there is a vast literature that tries to give a general and insightful definition of what a paradox *is* and to provide valuable classifications of all possible paradoxes. For instance, Sorensen (2003) is an analysis of paradoxes in philosophy which provides a proper definition of a paradox and tries to give a taxonomy of all possible paradoxes in philosophy, classifying them in terms of their roles, arguments, or structures. The task of saying what a paradox is, in general, challenging since there are many different paradoxes with many different features. For the sake of concreteness, I will take the definition that (Sainsbury, 1995, 1) gives:

> This is what I understand by a paradox: an apparently unacceptable conclusion derived by apparently acceptable reasoning from apparently acceptable premises. Appearances have to deceive since the acceptable cannot lead by acceptable steps to the unacceptable. So, generally, we have a choice: either the conclusion is not really unacceptable, or else the starting point, or the reasoning, has some non-obvious flaw.

According to this definition, paradoxes are situations that we have to consider very seriously and are of extreme importance because they can pinpoint false assumptions or help us realise that what we thought were unacceptable conclusions are only counterintuitive.

To understand which kind of paradoxes are black hole paradoxes, let me give several possible classifications of paradoxes in terms of their subjects, origins or structures and discuss the relevance of each classification for black hole paradoxes. One possible way to classify paradoxes is in terms of their subjects:

Ontological paradoxes, which are those that deal with real phenomena. Sometimes they are also called *real paradoxes*.

Semantic paradoxes, which are those that deal with the explanation of real phenomena. They can be of theoretical (explanatory) character or observational character.

This first classification is somehow tangential to black hole paradoxes and is more associated with one's attitude towards realism or antirealism about physical phenomena. For instance, if one is a realist about black holes in quantum gravity, i.e. they think that quantum black holes are real, black hole paradoxes can be categorised as ontological paradoxes since they deal with real phenomena. On the contrary, if one thinks that quantum black holes are just a possible description of a phenomenon that goes

beyond the scope of our contemporary theories, black hole paradoxes can be classified as semantic paradoxes. We will return to this issue and the ontology of black holes in quantum gravity later in this thesis.
Another classification of paradoxes has been proposed by Sainsbury (2009):

Scientific paradoxes, created by a scientific understanding of the world in which observations, experiments and calculations play a crucial role.

Philosophical paradoxes, originated by a philosophical understanding of the world that cannot be profitably incorporated into science.

In this respect, we can say safely that black hole paradoxes are *scientific paradoxes* since concrete scientific theories generate them. We will now focus on scientific paradoxes and try to give additional elements to distinguish between different paradoxes in physics. Another possible classification is in terms of the *cause* of the paradox:

Paradoxes of errors caused by a mistake that can be difficult to recognise. If the error is discovered, the paradox is solved. Thus, the essential characteristic of this type of paradox is that to solve it means to improve the understanding of the theory. However, the paradox is not fatal: the solution to the paradox does not change the paradigm but helps to understand the theory under consideration deeply.

Paradoxes of contradictions caused by a contradiction between different principles in a physical theory. To solve the paradox, one of the instances should be dropped. Thus, the essential characteristic of this type of paradox is that solving it means drastically changing the physical theory under study. Thus, it is fatal: the solution to the paradox changes the paradigm.

The paradoxes of contradictions are thus fundamental since they often appear in critical situations in which a new paradigmatic view in physics is formed. Indeed, they require relevant modification to our physical principles to be solved. Black hole paradoxes can be considered *paradoxes of contradictions*. Indeed, as we will discuss in the next chapters of this thesis, they originate as inconsistencies between well-accepted principles of our best physical theories, i.e. General Relativity and Quantum Mechanics.
Paradoxes can also be divided according to their genesis, i.e. whether they originate in observational or theoretical situations in physics:

Experimental paradoxes appear when experiments are performed, and the results of such experiments have no satisfactory explanation.

Theoretical paradoxes appear when a theoretical explanation of a physical phenomenon has not been precisely formulated.

Concerning theoretical paradoxes, one can think of a further classification:

Paradoxes that are thought experiments, i.e. paradoxes created as a result of theoretical speculations.

Paradoxes that are not thought experiments, i.e. paradoxes created by explaining a physical phenomena.

In this respect, we can safely say that black hole paradoxes are theoretical. Indeed, we cannot concretely perform experiments on real black holes in the universe. However, we will see along this thesis that some black hole paradoxes are thought experiments and others are not. Indeed, some paradoxes concern the features of Hawking radiation, which can be considered a physical phenomenon, and thus are not thought experiments. Some others, instead, concern theoretical speculations and are thus thought experiments.

Passing through some possible classifications of paradoxes in philosophy and physics, I tried to identify underlying features of black hole paradoxes. We have seen that they can be considered physical paradoxes of theoretical nature that concern contradictions between well-accepted physical principles. In the next section, we will provide examples of paradoxes in modern theoretical physics that share the same feature. We do this because we would like to analyse the role played for scientific progress by these paradoxes to understand which role black hole paradoxes can play in contemporary research.

2.2 Paradoxes in Modern Theoretical Physics

This section provides three examples of theoretical paradoxes originating from a contradiction between well-accepted physical principles. We will see how they shed light on the structure of more fundamental physical theories. This section's content must be understood as a conceptual reconstruction of the paradoxes under consideration rather than a historical one. The main goal is to (retrospectively) point out the real contradiction that the paradoxes pinpointed.

The first paradox I am going to analyse is the so-called *train paradox* (§2.2.1), which has been understood as pointing out contradictions between Newtonian Mechanics and Electromagnetism. The second paradox will be the so-called *elevator paradox* (§2.2.2), which can be understood as the clash between the Newtonian theory of gravitation and Einstein's basic ideas of Relativity. The third one is the so-called *EPR paradox* (§2.2.3), and it is a paradox concerning entanglement in Quantum Mechanics. In the section (§2.3), we will see how these three paradoxes contributed in different ways to scientific progress.

2.2.1 The Train Paradox

The train paradox can be understood as the clash of three well-established physical principles:

(i) **The Principle of Relativity**: the laws of nature are the same in different frames of reference.

(ii) **The Principle of the Constancy of the Speed of Light**: the speed of light in free space (vacuum) has the same value c in all inertial systems, i.e. the speed of light is a law of nature.

(iii) **Galilean Relativity**: the laws of mechanics are the same for all inertial observers, i.e. for observers that are standing still or moving at constant velocity v.

Each of these principles seems quite plausible on its own. However, as Einstein pointed out in Einstein (1905), they can't all be true simultaneously. To show the inconsistency between these three principles, Einstein developed the first paradox that we will discuss, a thought experiment.
Let us consider an observer, Alice, on a train which is moving with velocity v relative to another observer, Bob, standing still on the ground. Moreover, let us consider that Alice has a bulb and that light is spreading over the train according to Maxwell's equation, i.e. as an electromagnetic wave propagating with velocity c. If (i) is true, then for Alice and Bob, the laws of nature should be the same since the laws of nature should be the same in different reference frames. However, according to (iii), Bob should observe light propagating with velocity $c + v$, violating (ii). We thus have a paradoxical situation in which the three well-established principles (i), (ii) and (iii) conflict with each other. To

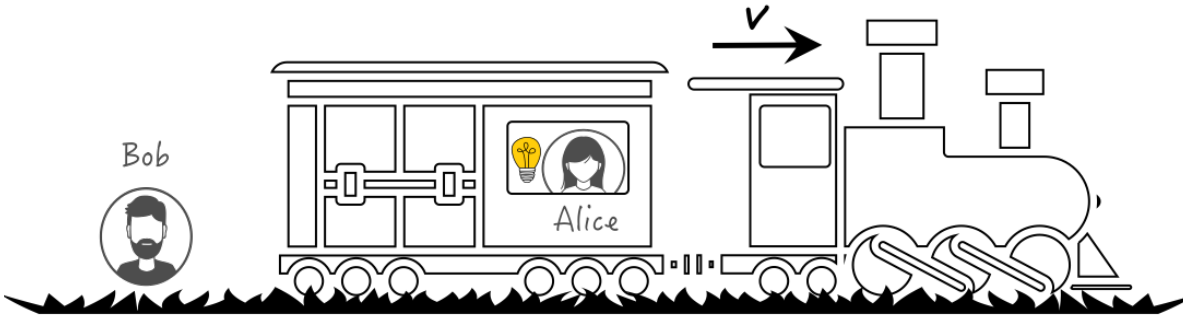


Figure 2.1: ...

solve the paradox, we should drop one of those.

Understanding which principle one should drop was not easy, and it never is for any paradox of contradiction, as we are going to see also for black hole paradoxes. Indeed, (i) seems a very general principle and thus a good requirement for any physical theory. In favour of dropping (ii), there was the idea that all Newtonian mechanics' laws were invariant under Galilean transformations, and thus (iii) was satisfied. Thus, dropping (ii) was a cheaper resolution to the paradox because dropping (iii) pointed out that Newtonian Mechanics and the Newtonian theory of gravitation had to be reformulated. On the contrary, it was already well-known that Maxwell equations were not invariant under Galilean transformation. Moreover, from Maxwell equations, one could understand that the speed of light was related to vacuum electromagnetic constants μ_0 and ϵ_0 and that it does not refer to any preferred frame or observer. In other words, Electrodynamics was not in conflict with dropping (iii) and was pointing to the direction of holding (ii). From this point of view, the paradox can thus be understood as the conflict between Newtonian's mechanic's theory and Maxwell's electromagnetism theory.

A resolution to the paradox meant to preserve Newtonian Mechanics was that we could drop (ii) and say that the speed of light, like any other entity, can differ depending on one's speed relative to it. But experiments designed to detect differences in the speed of light failed to do so Michelson and Morley (1887). Einstein proposed to drop (iii) and studied the consequences of such proposals. Giving up (iii) means that Newtonian Mechanics and Gravity cannot be correct, thus gaining electromagnetism but losing gravity. It turned out that the dropping of (iii) led to the theory of Special Relativity and a fundamental rethinking of our entire view of space and time. We have thus seen how the resolution of a contradiction paradox has guided research towards a new paradigmatic view.

2.2.2 The Elevator Paradox

The Principle of Relativity discussed in the previous section states that all laws of physics should be the same in different frames of reference. However, renouncing (iii), i.e. Galilean Relativity, Newton's theory of gravitation, which describes the orbits of planets around the Sun and the tides caused by the gravitational attraction of the Moon and the Sun with great accuracy, should be reformulated. This section aims to elucidate, formulating a second paradox, in which Newton's theory is incompatible with Relativity and how Einstein in Einstein (1915) managed to find a solution to this paradox. The key point of the discussion will be the relation between gravity and inertia. The resolution of the paradox will lead us very quickly to a geometric picture of gravity.

First of all, let me recall that in Newtonian physics, there are two *a priori* different concepts of mass:

- the *inertial mass* m_i can be understood as the resistance of a particle against acceleration. It appears, in particular, in the second law of Newtonian Mechanics:

$$\vec{F} = m_i \vec{a} \ . \tag{2.1}$$

- the *gravitational mass* m_g, which can be understood as the "gravitational charge" of a particle:

$$\vec{F}_g = -m_g \vec{\nabla} \phi \ . \tag{2.2}$$

As it is well-known, these conceptually different physical quantities are empirically equivalent. However, Newton's theory would be *in principle* consistent with $m_i \neq m_g$. It was an empirical fact that the gravitational and inertial masses must be the same. And indeed, Einstein was impressed by this equality and believed that it was not a coincidence but could tell us something deep about the nature of gravity. Moreover, note that this equality is a very peculiar feature of gravity: electromagnetism, for instance, does not have this feature since the equation for an electrically charged particle moving in an electrostatic field

$$m_i \ddot{x} = -q_e \vec{\nabla} \phi \ , \tag{2.3}$$

is compatible with any ratio q_e/m_i. The equality between the "charge", and the inertial mass is a very peculiar feature of gravity that is not shared by other forces.
Since they are conceptually different, let us work under the hypothesis that inertia and gravitation are two physically different concepts, even if the two masses are equal. And, under this assumption, let us then consider these two principles:

(i) **The Principle of Relativity**: the laws of nature are the same in different frames of reference.

(ii) **Gravity $\neq$ Inertia**: gravitation and inertia are conceptually different.

To see how these two principles conflict, Einstein highlighted a paradoxical situation by the so-called *elevator thought experiment*. Consider an observer, Bob, in a small elevator in outer space, as in the box (A) of figure 2.3. Bob will freely float since no forces are acting on him. Now assume to move to a constant accelerating frame, i.e. assume that somebody on the outside of the elevator pulls the elevator up with a constant acceleration, as in the box (B) of figure 2.3. Then, according to Newton's second law, Bob will be pressed to the elevator floor with constant force. Consider the same elevator brought into a constant gravitational field and fixed to the roof by a resistant cable, as in a box (C) of figure 2.3. Then again, Bob will be pressed to the floor with constant force. If the gravitational and inertial masses are equal, the force experienced by Bob by going to an accelerating frame and the force experienced because of the presence of a gravitational field should be equal. Thus, there is no experiment Bob can do to distinguish the effect of a gravitational field and a constant external acceleration. In other words, by principle (i), the laws of nature describing the physics in the box (B) and (C) should be the same. However, if (ii) holds, they are different since the first is an effect due to inertia, while the second is an effect due to gravitation. Now consider somebody cutting the cable of the elevator, as in the box (D) of figure 2.3. In this case, the elevator will freely fall downwards, and Bob will float in the elevator as if the elevator was in outer space (box (A) of figure 2.3). Thus, locally the effect of gravity can be eliminated by going to fall reference frames freely. In other words, if (i) is true, the laws of nature describing the physics in the box (A) and (D) should be the

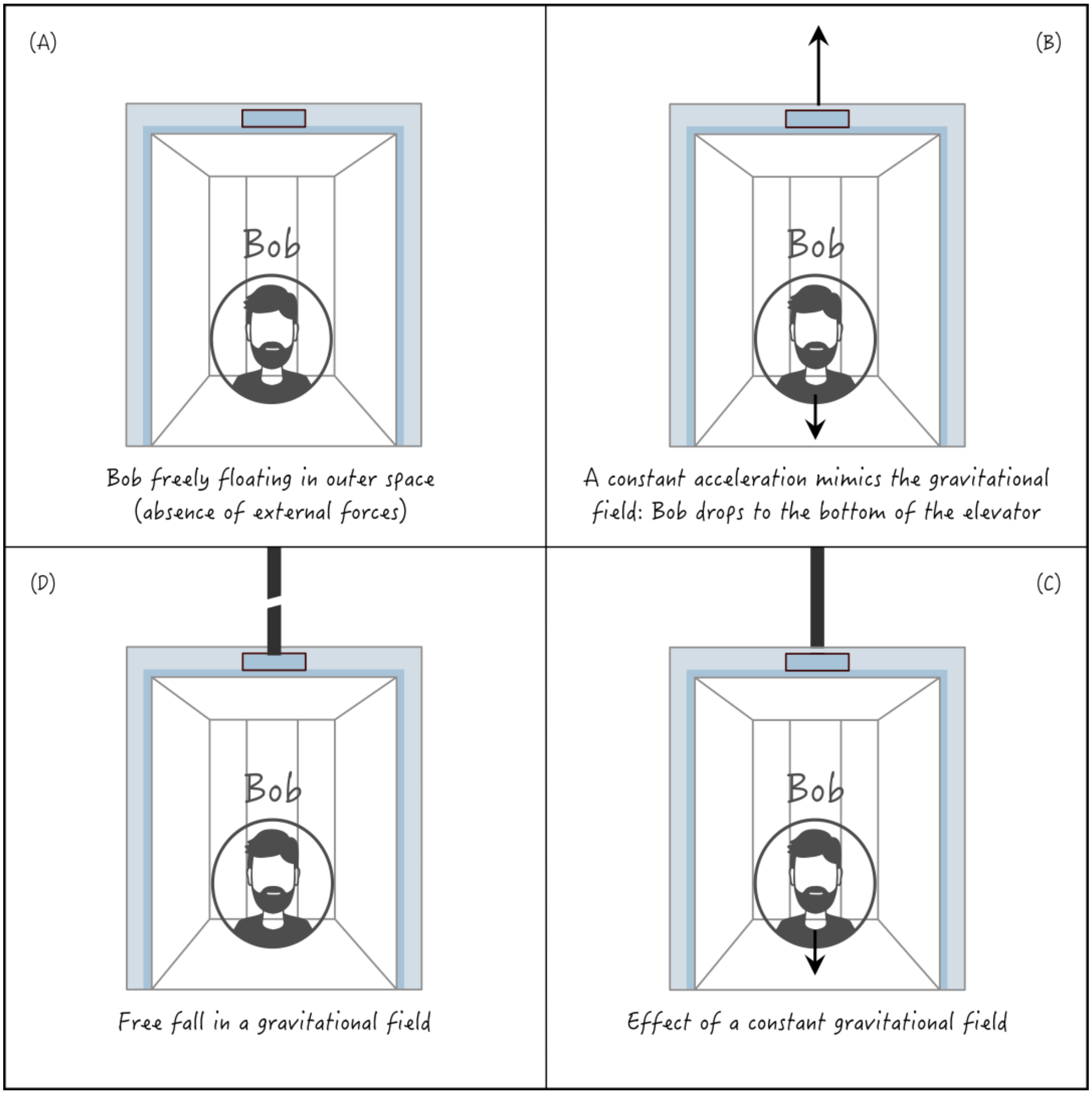


Figure 2.2: ...

same. Thus, the absence of inertial forces should be conceptually equal to the absence of gravitation, contra (ii).[2]

Let us consider the same experiment on a more mathematical note. We would have to perform a change in the reference frames to describe the physics in each box of figure 2.3. Let us consider the box (C) of figure 2.3. The following equation describes the

[2]Note that this version of the famous Einstein though the experiment was criticised by many philosophers afterwards (). This is because it does not consider tidal forces that would allow Bob to understand in which of the scenarios he is. However, I choose to have this version of the thought experiment for ease of exposition, and nothing in what follows hinges on this choice.

laws of physics inside the elevator:

$$m_i \ddot{z} = -m_g g \ . \tag{2.4}$$

If we were to describe the same physics on an accelerated reference frame and go to a freely falling coordinate system, we would need to perform the following coordinate transformation:

$$Z(z,t) = z + \frac{1}{2} g t^2 \ . \tag{2.5}$$

With this coordinate transformation, the physics in the new reference frame is described by the following equation:

$$m_i(\ddot{Z} - g) = -m_g g \quad \rightarrow \quad \ddot{Z} = (m_i - m_g) g = 0 \ , \tag{2.6}$$

where the last equality holds if the gravitational and inertial mass are taken to be the same. Thus, we see the contradiction between the two principles (i) and (ii). If the physics should be the same in all reference frames (i), gravity cannot be different from inertia since; otherwise, we would have that the freely falling experimenter will experience a force $\vec{F} = (m_i - m_g)\vec{g}$, which should not be the case if the principle of relativity hold.

Facing this paradoxical situation, we must drop one of the two principles. Einstein's choice was to drop (ii), i.e. considering gravity and inertia as the same thing. This choice was then not only dictated by experimental evidence, but the thought experiment elucidated that it was necessary if we would like the principle of relativity to hold. According to the legend, Einstein said this was his "happiest thought": his older theory of relativity, special relativity, holds locally, i.e. in small enough regions of spacetime. Einstein's reasoning was later formulated as the Einstein Equivalence Principle:[3]

> Gravitational and inertial effects are conceptually unified, representing the components of the same (possibly curved) compatible connection.

In other words, if you are in a small, enclosed space and you cannot see outside, you would not be able to tell whether you are in a gravitational field (such as the surface

[3] The precise statement of the Einstein Equivalence Principle has been taken by Read and Teh (2022). ...

of the Earth) or in a frame of reference that is accelerating uniformly (such as an elevator that is accelerating upwards). Starting from his equivalence principle, Einstein then developed the theory of General Relativity, in which the laws of physics locally are taken to be the one of Special Relativity. The equivalence principle thus forms the basis for the theory of General Relativity, which describes the behaviour of objects and the curvature of spacetime in the presence of matter and energy. In General Relativity, the gravitational force is not a force in the traditional sense but rather a manifestation of the curvature of spacetime caused by the presence of matter or energy.

Note that we have not mentioned Special Relativity until the end of this section. It is commonly thought that the conflict between Newtonian Gravitation and Relativity stands from the fact that the gravitational force in the Newtonian theory is an instantaneous force contra the principle that no signal can travel faster than the speed of light, a cornerstone of Special Relativity. This is, for sure, a paradoxical situation that has to be fixed, as argued in section §2.2.1. However, in this section, we saw that the conflict between Newtonian Gravitation and relativity stands from the fact that if we would like (i) to hold, we have to drop (ii), i.e. accept the fact that gravity and inertia are the very same thing. This has nothing to do with special relativity and the bound on signals imposed by the speed of light limit. And, indeed, one can also think of a non-relativistic theory of gravitation that is compatible with (i): the Newton-Cartan theory Cartan (1923).[4] Indeed, the upshot of the thought experiment was that we could eliminate gravity's effects locally. Einstein has decided to consider that locally the laws of physics are the one of Special Relativity, his preferred non-gravitational theory. However, nothing prevents us from thinking that, locally, the laws of physics are different from those of Special Relativity and are, for instance, the covariant generalisation of Newtonian Mechanics. Indeed, this choice would give rise to new non-relativistic theories of gravity compatible with (i) (Hartong and Obers, 2015). As we would expect from this reasoning, Read and Teh (2022) showed recently that these non-relativistic theories of gravity obey the Einstein equivalence principle. Of course, these theories would violate the speed of light limit, and they cannot be (most probably) the most fundamental theories we can build since they conflict with Special Relativity. However, they can be good gravitational theories, compatible with the principle of relativity, for describing low-velocity physics.

[4] For a modern take on Newton-Cartan theory and its generalisations see Christensen et al. (2014).

2.2.3 The EPR Paradox

The EPR paradox (Einstein et al., 1935), named after Albert Einstein, Boris Podolsky, and Nathan Rosen, is a thought experiment in quantum mechanics that aims to show the incompleteness of Quantum Mechanics. The EPR paradox is based on "entanglement," a property of specific pairs of particles that are created together and remain connected, even when separated by large distances. According to quantum mechanics, the properties of these entangled particles are correlated, and the state of one particle can affect the state of the other, even if a large distance separates them.
The EPR paradox emerges as the incompatibility of the conjunction of three principles:

(i) **Realism**: if without interacting with a given system, it is possible to predict with certainty (i.e. with probability one) the value of a particular physical quantity, this physical quantity corresponds to an objective property of the system, i.e. a property independent of any observers.

(ii) **Locality**: given two physical systems and supposing that during a specific time interval Δt they remain isolated from each other (which is achieved in Special Relativity if the distance between them is greater than $c\Delta t$, i.e. if they are space-like related), then the evolution of the physical properties of one of them during this time interval cannot be influenced by operations performed on the other.

(iii) **Completeness**: if any element of the objective reality has a counterpart among the concepts used by a given theory, then the theory is said to be complete.

Einstein et al. (1935), using a thought experiment, showed that there is a clash between these well-established principles and that we have thus to drop one of them. This section will review the Bohm version of the EPR paradox.
Consider a bulb emitting a large number of pairs of spin 1/2 particles, and consider the pairs of spin 1/2 particles to be in a maximally entangled state, such as *the singlet state*:

$$|\psi\rangle = \frac{1}{\sqrt{2}} \left(|\uparrow\downarrow\rangle - |\downarrow\uparrow\rangle \right) \, , \tag{2.7}$$

where the first entry in the ket refers to the electron emitted to the left of the bulb, while the second entry refers to the electron emitted to the right of the bulb. At some time t_0, the supports of the spatial wave functions of the two electrons are at a distance

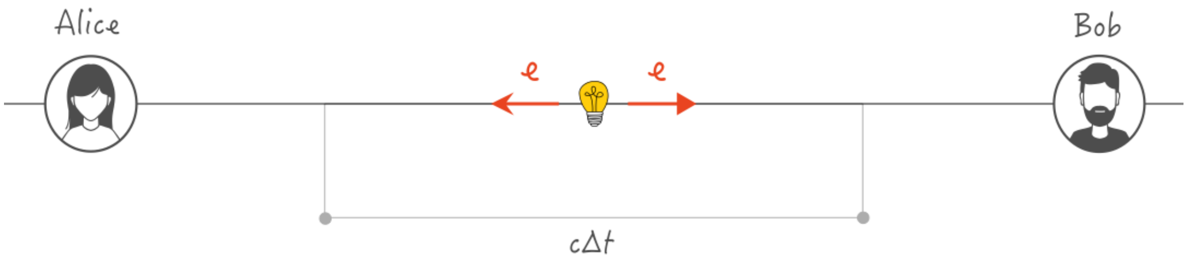


Figure 2.3: ...

d such that they cannot interact with each other without violating (ii). Now let us consider, as in figure 2.3, two experimenters, Alice and Bob, collecting the electrons emitted by the bulb. At $t_1 > t_0$, Alice can measure the spin along the z axes on her electrons, and she will observe half of the time spin up and half of the time spin down. If she, for instance, finds spin up on her electron, the system's state will collapse by the Von Neumann projection postulate to $|\uparrow\downarrow\rangle$. Thus, at $t_1 > t_0$, we are certain that Bob would find spin down along the z axes on his electrons. Moreover, if (ii) holds, we also know that the measurements performed by Alice cannot perturb Bob's electron.[5] Thus, since Alice's measurement cannot produce any changes in Bob's electron after t_0, Bob's electron should have had a definite spin also at time $t_0 < t < t_1$. Thus that specific pair of electrons would be in the singlet state before t_0 and in the state $|\uparrow\downarrow\rangle$ after t_0.

The experiment can be repeated many times on each pair of electrons, and, half of the time, the state will collapse to $|\uparrow\downarrow\rangle$ and half of the time to $|\downarrow\uparrow\rangle$. Thus the state at times $t_0 < t < t_1$ would be described by a mixed state:

$$\rho = \frac{1}{2} |\uparrow\downarrow\rangle \langle\uparrow\downarrow| + \frac{1}{2} |\downarrow\uparrow\rangle \langle\downarrow\uparrow| \ . \tag{2.8}$$

Here we already find an inconsistency: for times $t < t_1$, no measure has been performed yet, and thus we do not expect any violation of the unitary evolution of the system. However, we showed that we have transitioned from a pure state (2.7) to a mixed state (2.8), thus violating unitarity. Moreover, the unitary evolution of the state (2.7) and the state (2.8) will produce two different predictions on the mean values of different operators, such as spin x or spin y.

[5] Modulo Separability, which is implicitly assumed in Locality by Einstein et al. (1935).

We can see now the clash between the three principles:

- The conjunction of (i) and (ii) leads to a violation of a fundamental principle of quantum mechanics. Thus (iii) cannot hold, i.e. quantum mechanics is not a complete theory.
- (iii) holds, i.e. quantum mechanics is complete. Thus either (i) or (ii) cannot hold. Otherwise, we fall into a contradiction.

The paradox could be thus resolved by dropping one of the three principles:

Drop (i): one could renounce Realism by saying that even knowing with certainty the value that would be obtained by measuring an observable without disturbing the system, this value cannot be attributed to the character of an objective physical property independent of the measurement. In particular, we cannot extrapolate the information before the measurement, even if we know the system's dynamics. In this way, we cannot "bring back" the result obtained in t_1, and there is no contradiction for $t_0 < t < t_1$.

Drop (ii): one could renounce Locality by saying that a measurement can instantly change the wave function at arbitrarily large distances. Therefore, the information cannot be extended backwards even if the systems are at space-like distances, given that they could, in principle, even in this case, influence each other. Even if this contrasts with the theory of Special Relativity, we will see in the following chapters of this thesis that entanglement and special relativity are compatible. However, this will become an issue when gravitational interactions are considered, and we will have to rely on a different understanding of Locality.

Drop (iii): one could renounce Causality by saying that the spins of Alice and Bob were already defined before Alice's measurement but were not knowable via (2.7), which therefore constituted incomplete information of the system. Thus (2.7) only gives probabilistic information as the mean of "hidden" variables, which constitute the "true" real variables. An analogy is a temperature in Statistical Mechanics, which is simply an average over a huge number of particles whose states are not individually measured.

Option three of this list has been ruled out by Bell (Bell and Bell, 2004), showing that quantum mechanics and local hidden variable theories are incompatible. Therefore, it is impossible to drop (iii) since the alternative theories are not valid, and it follows that quantum mechanics is a complete theory. The only possible solutions to the EPR paradox are thus the ones dropping either Realism or Locality, as understood by EPR.

2.3 On the role of Paradoxes for Scientific Progress

In the previous section, we reviewed three famous paradoxes of the past century and have seen how they have been of extreme importance for developing our best physical theories: Relativity (broadly speaking) and Quantum Mechanics. In this section, I would like to analyse the specific role played by these paradoxes for scientific progress and which lesson we could learn on the significance of black hole paradoxes for contemporary research.

The first paradox, discussed in §2.2.1, was put into use directly on theory construction. In other words, we can think of Einstein's theory of Special Relativity as having its origins in the train paradox. Indeed, quoting Einstein (2013):

> If a ray of light is sent along the embankment, we see from the above that the tip of the ray will be transmitted with the velocity c relative to the embankment. Let us suppose that our railway carriage is again travelling along the railway lines with the velocity v and that its direction is the same as that of the ray of light, but its velocity is much less. Let us inquire about the velocity of propagation of the ray of light relative to the carriage. We can here apply the consideration of the previous section since the ray of light plays the part of the man walking along relatively to the carriage. The velocity W of the man relative to the embankment is replaced by the velocity of light relative to the embankment. w is the required velocity of light with respect to the carriage, and we have
>
> $$W = c - v \ . \tag{2.9}$$
>
> The propagation velocity of a ray of light relative to the carriage thus comes out smaller than c.

Thus, Einstein presents the train paradox as guiding the development of Special Relativity. Indeed, facing the inconsistency between the three principles (Relativity, Constancy of the Speed of Light and Galilean Relativity), Einstein decided to drop the latter and thus searched for a new form of transformation suitable to be consistent with both Relativity and Constancy of Speed of Light. In this way, he solved the paradox and put the basis to formally derive the theory of Special Relativity that has revolutionised our conception of space and time.

The second paradox we analysed, i.e. the elevator paradox of section §2.2.2, can be seen from a different perspective about scientific progress. Indeed, we cannot say that the elevator paradox contributed to the development of General Relativity as it was to the train paradox with Special Relativity. Indeed, Einstein proposed the elevator paradox later on. The experiment that guided him should be that of the two softballs in an otherwise empty universe, one spinning and becoming ovoid, the other, not. However, we can attribute another important role to the elevator paradox. Reconstructing the paradox following the way proposed in section §2.2.2, we have been able to disentangle two different inconsistencies: the inconsistency between the Galilean character of Newton's Gravitation and the Constancy of the Speed of Light and the inconsistency between the Principle of Relativity and the conceptual separation between gravity and inertia. Having disentangled these two inconsistencies, we have seen that the elevator paradox, and thus the foundation of General Relativity, involves the second, not the first. Indeed, as we were also mentioning at the end of the section §2.2.2, the solution of the elevator paradox proposed by Einstien, i.e. the Einstein equivalence principle, is also shared by covariant formulations of non-relativistic theories of gravity. Thus, we can say that the elevator paradox not only motivated the origin of General Relativity, but a more careful look at the paradox also motivated new theories of gravitation with a non-relativistic character. In other words, a more careful analysis of the elevator paradox helped researchers to generalise General Relativity discovering a broader family of gravitational theories, which differs for the local group of spacetime taken into account. These new theories of gravitation could not be fundamental descriptions of our world because they are still in conflict with the Constancy of the Speed of Light since they are non-relativistic (or ultra-relativistic). Indeed, only General Relativity can solve both the inconsistencies mentioned above simultaneously. However, they are coarse-grained descriptions which can describe our world at low velocities. And indeed, it turned out that these theories can be fundamental in many different physical situations (Mukohyama, 2010; Son, 2013; Hartong et al., 2016). On a more historical note, it has been shown by Hansen et al. (2019) and Hansen et al. (2020) that also these new non-relativistic theories of gravitation pass the so-called *classical tests of General Relativity*, i.e. the "anomalous" precession of the perihelion of Mercury, the bending of light in gravitational fields, and the gravitational redshift. This means that the classical tests of General Relativity are not testing General Relativity but rather the new family of gravitational theories that share only one common feature: the Equivalence Principle. On this note, James Read (202x) argues that the classical tests of General Relativity are thus not testing General Relativity but the Equivalence Principle itself. The only feature that can distinguish between General Relativity and the non-relativistic theories of gravitation is the presence of gravitational waves. Thus, the first real test of the validity of General Relativity is the detection of gravitational waves, which has been realised only in Abbott et al. (2016), approximately one hundred years later than General Relativity. We have thus seen that the elevator paradox has not only been important for the development of General Relativity, but a closer look at the paradox has been able to help researchers to develop different theories of gravitation, which shed light on the empirical status of General Relativity, and can also help us better to understand the ontology of gravitational theories in general.

The EPR paradox, instead, has been important for a very different reason. Indeed, the EPR paradox has yet to lead to a new theory that could solve the inconsistency, as Einstein, Podolsky, and Rosen would have hoped. Rather, it helped us understand more deeply the theory we had, i.e. Quantum Mechanics. After the empirical tests of Bell's inequality, we can confidently say Quantum Mechanics is a complete theory of the microscopic world at the non-relativistic and non-gravitational levels. However, we cannot say that we have understood what Quantum Mechanics says about the real world. Indeed, the philosophy of Quantum Mechanics is a very active field of research. The EPR paradox has been an essential tool for understanding various features of quantum mechanics, both from interpretative and metaphysical levels. In other words, the EPR paradox did not have a role in theory construction, but it helped us better understand and highlight the quantum theory's crucial feature.

We have thus seen that in the history of modern theoretical physics, paradoxes have played three distinct roles:

Construction of a theory: paradoxes can help to develop more fundamental theories.

Development of a theory: paradoxes can help to disentangle different features of a given theory, thus leading to further developments within physics, philosophy of physics, and metaphysics.

Understanding of a theory: paradoxes can help to deeply understand the consequences of our physical theories, thus providing a valuable tool for philosophical and metaphysical considerations.

In the next section, we will see how black hole paradoxes can play similar roles in contemporary physics.

2.4 Paradoxes and Quantum Gravity

The development of a coherent theory of quantum gravity has been the most studied and important problem in theoretical physics over the last 50 years. Quantum gravity should be a theory which can unify General Relativity, which is a description of the gravitational force in terms of the curvature of spacetime, with quantum mechanics, which is a description of the behaviour of matter on a microscopic level. In this context, black holes and, in particular, their quantum properties are fundamental since they have been the source of progress and new ideas in the search for a possible theory of quantum gravity. This is because it is one of the few systems for which we have to consider relativistic effects and quantum effects at the same time. To see why let's start with the observation that general relativity is crucial for describing black holes. Indeed, black holes, being regions where the force of gravity, and therefore the curvature of space-time, are such as not to allow even light to escape, are objects described by General Relativity. At the same time, quantum theory also plays a crucial role in describing black holes. We can realise this by thinking about the singularity hidden inside the black hole or that point in spacetime where the curvature of spacetime becomes infinite, and General Relativity ceases to be usable. The singularity, and the region immediately surrounding it, is a microscopic region of spacetime; therefore, it is reasonable to expect quantum theory to be important in its description. At the same time, being (the region adjacent to) the singularity part of the black hole and being a region of high curvature, General Relativity will also be crucial in its description. Therefore, a theory of quantum gravity is necessary to describe a black hole fully. And consequently, a better understanding of the structure of black holes, particularly their quantum properties, is crucial for developing a theory of quantum gravity and, therefore, for contemporary theoretical physics.

Since the pivotal paper on Hawking radiation (Hawking, 1975), it was already well understood that when one tries to describe quantum properties of black holes by putting together, even at an approximate level, quantum mechanics and general relativity, they encounter inconsistency giving rise to several paradoxes. I'm particularly interested in black hole paradoxes since their resolution could be an open window into the realm of quantum gravity. Indeed, these paradoxes have important repercussions on a myriad of questions that touch the deepest foundations of general relativity and quantum mechanics: think, for example, of the question of whether quantum gravity is a unitary theory or not, directly affected by the so-called *information loss paradox*. Or to the questions that arise concerning the nature of a black hole's microscopic degrees of freedom, i.e. the elementary constituents that would hypothetically compose the black hole, where it is not clear whether thermodynamic quantities such as the entropy of a black hole can be interpreted starting from the statistical-mechanical properties of the microscopic states of the black hole, as happens for all other thermodynamic quantities through statistical mechanics. This issue is doubly linked to the so-called *Page paradox*. Moreover, in recent years, the paradox that has most attracted the attention of the community of theoretical physicists is the so-called *firewall paradox*, a paradox that gives rise to even more radical and fundamental questions than those raised by the other two paradoxes mentioned above: the firewall paradox seems to suggest the non-existence of the interior of the black hole, which, to ensure the possibility of a coherent theory of these cosmic objects, should instead be replaced, at least according to Almheiri et al. (2013), with a wall of flames.

It is essential to analyse the conceptual foundations of these paradoxes and their resolution, following the three main paradigms we developed in the previous section. First, by exploring the space of the logical resolution to these paradoxes, one could help theoretical physicists construct a theory of quantum gravity. Second, by deeply analysing the conceptual structure of these paradoxes, one might recognise implicit assumptions and disentangle the main physical principles that come into play, helping working physicists develop different resolutions and alternative theories. Lastly, and perhaps even more importantly, the philosophical analyses of these paradoxes can help us understand the true nature of quantum black holes, being able to put forward a possible ontology of these objects.

It is worth mentioning that philosophers have not ignored the importance of studying the conceptual structure of black holes and their paradoxes. See, for example, to cite some works dedicated to this topic, Wallace (2020a), and Belot et al. (1999b). However, both of the articles mentioned limit themselves to dealing with the paradox of the loss of information and Page's paradox of time, which both revolve, as we shall see, around the idea originally discussed by Hawking that the physics of black holes may be non-unitary, i.e. incompatible with one of the fundamental principles of quantum mechanics, and on the various possibilities for avoiding this conclusion. On the other hand, the firewall paradox plays a central role in understanding what it means to describe the inside of a black hole and when this is possible. My goal, therefore, as a philosopher of physics, is to study the conceptual foundations of the firewall paradox and explore which conceptual strategies and modifications to the basic structure of general relativity and quantum mechanics it is possible to solve it. Therefore, we do not aim to develop new physics but rather to extract the fundamental lessons that the physics already developed by physicists can teach us about the structure of black holes. Indeed, We have seen in the previous section that, to formulate a paradox formally, we need to be able to write down the conceptual structure of a given theory. In other words, one has to understand the basic principles of a given theory to see which ones enter into conflict and to highlight hidden principles or assumptions. This is not always an easy task and is often overlooked by physicists nowadays. However, it is essential to study the consequence of paradoxes to provide such a conceptual analysis. This is one way in which the philosophy of physics can contribute to cutting-edge research in modern theoretical physics: elucidate the conceptual structure of a theory and understand if some of the principles give rise to paradoxes. And this is one of the most important tasks that the philosophy of physics can accomplish, and it will be the task of the next chapters of my thesis.

In this light, paradoxes are not bad situations to face but rather are the most exciting and generative conditions in which one can be. They cannot, for sure, substitute experiments. Experiments remain the primary tool to distinguish between true and untrue statements. However, paradoxes can be a good starting point for any theorists lacking experiments. They can be keystones to new theories, with many unexpected and exciting advances.

Chapter 3

Black Hole Paradoxes

This chapter presents some black hole paradoxes that have been a central topic of discussion among theoretical physicists over the last fifty years. In particular, we will focus on the developments that physicists have made in the past ten years, which, as we will see, have been quite outstanding. In this chapter, however, we will limit ourselves to presenting the physics of the paradoxes, i.e. introducing the theories and the physical explanation of why the paradoxes arise. The content of the following chapter will analyse the paradox's details, giving them a solid conceptual ground and presenting the possible solutions. We will, indeed, elucidate the conceptual structure of each paradox presented in this chapter and understand the principles that enter into conflict in these paradoxes, underlying implicit assumptions. But for the time being, we will give an (as much as possible) self-contained review of the physics of contemporary black hole paradoxes.

All black hole paradoxes derive from the attempt to describe the evaporation process of a black hole precisely, and it is, therefore, crucial to understand the characteristics of this process. The evaporation of black holes is a phenomenon described for the first time by the British physicist Stephen Hawking in a series of articles which have now become real classics (Hawking, 1975, 1976). The black hole evaporation is based on the derivation of the so-called *Hawking radiation*, i.e. the radiation emitted by the black hole. By analogy with the case of a boiling pot of water, we can imagine Hawking radiation as the vapour emitted by an evaporating system. The description of black holes as evaporating systems is perhaps the most important theoretical result over the past fifty years of the research attempting to combine Gravity and Quantum Mechanics. However, there is no direct experimental evidence for Hawking radiation to date. Indeed, for large black holes, such as those in our galaxy, the temperature of the Hawking radiation, according to Hawking (1976)'s results, would be lower than the temperature of the cosmic background radiation and therefore impossible to detect. Despite this, almost all physicists working in this area believe that the emission of Hawking radiation by a black hole should be taken for granted. The arguments usually put forward by physicists in favour of Hawking radiation are the following:

The first argument comes from the analogy between black holes and ordinary thermodynamic systems. To fully exploit this analogy, it is natural to expect black holes to emit radiation as ordinary thermodynamic objects do. Otherwise, black holes would have temperature and entropies like ordinary thermodynamic systems, but they wouldn't be able to interact thermally with their environment. Many physicists, therefore, believe that Hawking's calculation can be thought to be the confirmation of the analogy between black holes and ordinary thermodynamic systems.

The second argument is that Hawking's results themselves can be derived from diverse theoretical frameworks and with the most disparate techniques.[1] This convergence of different theoretical approaches suggests that there is something underlying physics; at least, this is how the story goes.

The third argument concern recent experiments on the so-called *analogous systems*, i.e. physical systems whose theoretical description is analogous to that of a black hole, but which can be realised in the laboratory, such as sonic black holes (Unruh, 1981; Visser, 1998). These new experiments have shown that these analogue systems do indeed evaporate. However, this experimental confirmation is very indirect, and it is not clear to what extent these experiments verify the existence of Hawking radiation.[2]

Although the epistemic status of Hawking radiation is an exciting and important topic, which philosophers should take into consideration, such questions go beyond the scope of this thesis. Indeed, if Hawking radiation did not exist, all modern theories of black holes would be false, and the paradoxes we are considering would be meaningless. However, since there are no adequate alternative theories, we will set aside these concerns

[1] Although see Gryb et al. (2021) for criticisms of how much different computations of Hawking radiation support such arguments.

[2] On this point, there is an interesting philosophical debate (Crowther et al., 2021; Dardashti et al., 2019).

and assume, with the majority of the theoretical physics community, that black holes radiate and evaporate and that this process is described, at least as a first approximation, by the calculations made by Hawking in Hawking (1976).

Before the publication of Almheiri et al. (2013), the two main black hole paradoxes were considered to be the *paradox of information loss*, initially introduced by Hawking (1976), which we will analyse in §3.1, and *Page's paradox of time*, initially introduced by Page (1993), which we will study in §3.2. We will review these two paradoxes before moving on to the firewall or AMPS paradox (§5.1) and to the black hole interior or AMPSS paradox (§5.2), which are the real focus of my thesis. As we will see, the first three paradoxes concern the black hole evaporation process, and the theoretical framework considered to derive them is Semiclassical Gravity. However, the last paradox we are going to consider, i.e. the AMPSS paradox, has been formulated using techniques from the holographic principle.

The chapter aims to be self-contained, thus providing a valuable introduction to black hole paradoxes for any philosophers of physics. For this reason, in any of the following sections, there will be an in-depth study of the relevant physical mechanism that generates the various paradoxes. In the first section (§3.1), devoted to the black hole information loss paradox, there will be a specific focus on Hawking radiation and the mechanism of black hole evaporation (§3.1.1). In the second section (§3.2), before presenting the Page time paradox, we will review some essential features of black hole thermodynamics quickly (§3.2.1). In the third section (§3.3), we are going to list the desiderata for a physical description of a black hole in quantum gravity proposed by Maldacena and Susskind (2013) (§3.3.1), which generates the AMPS paradox. Finally, in the fourth section (§3.4), we will review the holographic interior reconstruction techniques (§3.4.1,§3.4.2) leading to the AMPSS paradox.

3.1 Black Hole Information Loss Paradox

The black hole information loss paradox is a puzzle about information during the black hole evaporation process. In a nutshell, according to the principles of Quantum Mechanics, information about a physical system can never be completely destroyed. However, according to the theory of General Relativity, anything that falls into a black hole is either swallowed up by the singularity at the centre of the black hole or emitted as Hawking radiation. Hawking showed in Hawking (1976) that the Hawking radiation could not carry any information. Thus, we have a paradox, as it seems that the information contained in objects that fall into a black hole is being destroyed, contrary to the principles of Quantum Mechanics. In this section, we will first review Hawking radiation and the mechanism of black hole evaporation (§3.1.1), and then we will introduce the black hole information loss paradox (§3.1.2).

3.1.1 Hawking Radiation and Black Hole Evaporation

To describe Hawking radiation and, consequently, black hole evaporation, it is not enough to consider General Relativity only. Indeed, classical black holes do not radiate and thus do not evaporate. Therefore, we need to consider the possible quantum effects in the black hole's vicinity. In the absence of a fundamental theory of quantum gravity, there appear to be three main ways to get a first approximation to these effects:[3]

A first approach, which is the least sensitive to quantum effects, is to take quantum fields propagating in curved non-dynamical spacetime, i.e. the study of the so-called *quantum field theory in curved spacetime*. Since, in this approach, spacetime and the quantum field are treated as totally independent systems, all we can describe are the effects on the dynamics of a quantum field due to the curvature of spacetime. This approach underlies Hawking (1975)'s original radiation calculations.

A more sophisticated approach to the interaction between spacetime and quantum mechanics is the so-called semiclassical gravity, i.e. the theory obtained by replacing matter stress-energy tensor $T_{\mu\nu}$ in Einstein's equations with its expectation value $\langle T_{\mu\nu} \rangle$, obtained by treating $T_{\mu\nu}$ as the energy-momentum tensor of quantum fields living in the surrounding of the black hole. Since, in this approach, we incorporate quantum fields into Einstein's equations, we also consider the gravitational effects generated by the quantum field on spacetime, particularly the so-called *back-reaction* of the quantum field on the gravitational field. However, spacetime in semiclassical gravity is still treated as a purely classical entity, $g_{\mu\nu}$ still being the metric of general relativity.

Finally, the approach that comes closest to taking into account all the quantum effects of spacetime, at least in the absence of a complete theory of quantum gravity, is

[3] See Wallace (2022) for a technical and philosophical introduction to the second and third approaches. See Witten (2020) for a detailed discussion of the former.

the so-called *perturbative quantum gravity*, i.e. the theory obtained studying the path integral on the possible configurations of the gravitational field, order by order in perturbation theory. This corresponds to treating the gravitational field as an ordinary quantum field, like those of the Standard Model, and therefore taking into account its quantum properties already visible at low energies. This approach summarises the two previous approaches and allows us to verify the unitarity of the quantum evolution of a black hole, as we shall see later on in this thesis.

For the content of this section, it will be sufficient to focus our attention on the first two approaches, thus dealing with ordinary quantum fields in classical spacetime, at most taking into account their back-reaction on spacetime.
An essential feature of quantum fields is that small fluctuations in their vacuum state lead to the creation of particle/antiparticle pairs, which typically annihilate almost immediately. This phenomenon, characteristic of quantum fields, is crucial for under-

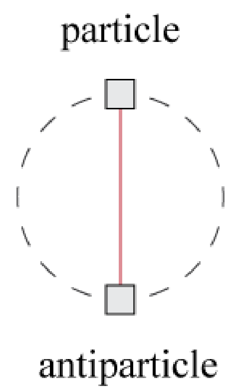


Figure 3.1: In the figure, the dotted line represents the propagation of a virtual particle or antiparticle, the boxes represent the particles/antiparticles, and the red line represents the entanglement between them. We use these conventions hereafter.

standing Hawking radiation. Indeed, let us imagine that the process of creation and destruction of a particle/antiparticle pair takes place across the horizon of a black hole. The particle or antiparticle created just inside the event horizon will fall towards the black hole's singularity. In contrast, the other, which lies just outside the black hole, will be free to escape from the black hole itself. Therefore, the two particles cannot annihilate, although they have been created by vacuum fluctuation. The particles escaping the black hole will be the constituents of Hawking radiation.[4] Starting from

[4]Keep in mind that this cartoonish explanation has not to be taken seriously and can be thought to be a thought experiment.

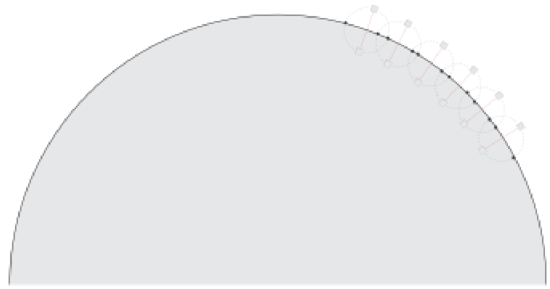


Figure 3.2: Vacuum polarisation across the horizon of a black hole.

this simple picture and considering properties of vacuum fluctuations and black holes in General Relativity, we can specify some essential features of the radiation emitted by the black hole with this mechanism and its consequences.

The density matrix of the Hawking radiation is thermal.

Since the two particles are in an entangled state,[5] we also know that the reduced density matrix of the external particle, which represents a book-keeping device for all the properties of the particle measurable by an external observer, will be thermal. Indeed, given a generic entangled state, the reduced density matrix of its subsystems will be thermal.[6] This thermal nature corresponds to the uncertainty that an observer who only had access to that density matrix of the radiation would have on the state of the composite system, i.e. black hole plus radiation. The uncertainty expresses that he cannot determine the correlations between the two subsystems, having access to only one.

To understand why we speak of *thermal* properties of the reduced density matrix, think about the interpretation of thermodynamic quantities that comes from statistical mechanics, such as entropy in terms of the uncertainty on the microscopic state of a given system. The reduced density matrix represents, as we said above, the uncertainty that an observer has on the properties of a particular quantum subsystem. Moreover, we can say that this uncertainty is due to the system itself, in the sense that it is not determined by the properties it possesses, except in a

[5]This is a feature of vacuum fluctuations: the particle/antiparticle pair is typically in a singlet state.

[6]To understand this point, it is useful to observe that the fact that having entanglement entropy greater than 0 corresponds to having a reduced density matrix with a thermal component. Maximum entanglement would correspond to completely thermally reduced density matrices, as in the case of subsystems of maximally entangled states such as those of vacuum fluctuations.

probabilistic sense. Given the correspondence between uncertainty and thermodynamics that we mentioned before, it is immediate to see why we speak of the *thermal* density matrix. However, it is useful to underline an essential difference between the quantum case and the relationship between statistical mechanics and thermodynamics. Since the uncertainty in the quantum case is a consequence of the indeterminacy of the properties of the quantum system itself, we cannot interpret this uncertainty in terms of the ignorance of an agent about the microscopic state of a system. In other words, we cannot give an epistemic interpretation to this indeterminacy, as instead is done in the case of statistical mechanics.

The Hawking radiation allows the black hole to be in thermal contact with its surroundings.

Back to our pair of particles straddling the hole horizon black hole, if we calculate in particular the entanglement entropy associated with the particle moving away from the horizon towards the outside of the black hole, we will obtain that its entanglement entropy and its temperature correspond to those of the black hole

$$S_{BH} = \frac{A_{hor}}{4G_N} \qquad , \qquad T = \frac{\kappa}{2\pi} \ . \tag{3.1}$$

Not only that. Since the two particles are a particle-antiparticle pair, the one that remained inside the black hole and fell towards the singularity will have energy, and therefore temperature, opposite to the one that escaped outside the black hole. Consequently, the black hole emits a particle that increases the temperature of the outside environment and engulfs one that lowers its own. This basic mechanism expresses nothing more than the fact that the black hole exchanges energy and heat with its surrounding environment or, in other words, is in thermal contact with it. We have thus obtained a mechanism for putting the black hole in thermal contact with the external environment, thus guaranteeing that we can treat the black hole as an actual thermodynamic system. We will call this mechanism the Hawking mechanism and the radiation emitted by the black hole Hawking radiation.[7]

The black hole is evaporating.

[7] From the name of the physicist Stephen Hawking who was the first, in his fundamental article Hawking (1975), to demonstrate the possibility of these phenomena mathematically.

In addition to being in thermal contact with its environment, we can also see that the Hawking mechanism ensures the evaporation of the black hole. Formally, this is demonstrated in the context of semiclassical gravity by calculating the back-reaction of the radiation on the geometry of the black hole. Back-reaction means the change caused by the propagation of the particles (from now on, however, we will call them quanta) of radiation on the metric itself. Intuitively, we can think of the evaporation process in the following way: we can consider that since the black hole emits energy via radiation, in the absence of an energy flow, to compensate for this loss, its energy will decrease. Since the energy expresses its mass (due to Einstein's relativity formula), its mass will also decrease. Equivalently, we can think of this process from the point of view of entropy exchange. The radiation has the effect of decreasing the entropy of the black hole. Furthermore, since the black hole's entropy is related to its area, its area decreases over time. In both senses, we can see that the black hole loses masses, and its area decreases over time, i.e. the black hole evaporates. Note that this drop in entropy and area does not conflict with the second law of thermodynamics. Indeed, the second law of thermodynamics states that an isolated system's entropy (or, in the black hole case, the area) cannot decrease. However, the black hole is no longer isolated in the scenario we are considering, as it is in thermal contact with its environment. If anything, the entropy of the black hole system, together with its surroundings, cannot decrease. The entropy, and therefore the area of the black hole, can instead, in this context, decrease without violating the second law of thermodynamics.

These are Hawking's crucial results on the evaporation of black holes. However, we will now see that these essential results generate several paradoxes in the physics of black holes.

3.1.2 Black Hole Information Loss Paradox

The black hole information loss paradox concerns the non-unitarity of the complete evaporation of a black hole.[8] To see why the features of the Hawking radiation mentioned at the end of the previous section are in contrast with the unitarity of the complete evaporation of the black hole, consider an evaporating black hole as in Figure 3.3.

[8] See Belot et al. (1999a).

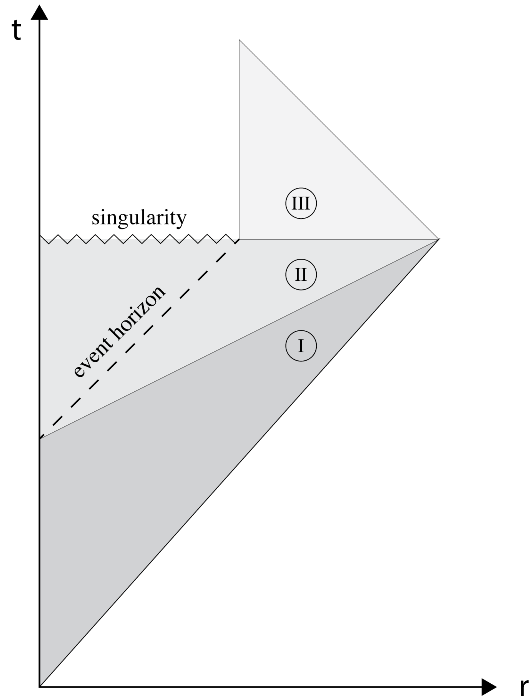


Figure 3.3: Spacetime diagram of an evaporating black hole.

In this Penrose diagram, we can distinguish three regions: (I) the pre-formation region, i.e. the collapse of a star into a black hole; (II) the evaporation region, i.e. the region in which the black hole starts evaporating due to the emission of Hawking quanta; (III) the post-evaporation region, i.e. the region in which the black hole has evaporated. As we were saying in the previous section, Hawking showed that the quantum state outside the black hole in the region (II) is in a mixed state, i.e. that the state of the radiation is perfectly thermal Hawking (1976). In other words, the quantum state of Hawking radiation is determined only by the macroscopic properties of the evaporating black hole, i.e. its mass, charge and angular momentum. In turn, Hawking radiation being exactly thermal also means that we cannot retrodict the state of the star collapsing into the black hole. The information encoded in this state should then be stored elsewhere, and the only possibility is the interior of the black hole.[9] Nevertheless, since the black hole eventually completely evaporates, the information can be stored in the interior of the black hole only region (II). Indeed, there are no slices in region (III) from which we

[9]Understanding the meaning of information in BHIP is certainly important. However, in this thesis, we are most interested in the AMPS and AMPSS paradox. Thus, for the time being, we bracket this issue since nothing of what we will say hinges on it.

can retrodict the state of the collapsing star. Thus, there cannot be unitary evolution from the region (I)+(II) to region (III). This is the main result of Hawking (1976)'s work: by considering quantum fields on a curved black hole background, it is possible to derive that the consequent black hole evaporation has to be non-unitary. We shall see in the next chapters that there are possible ways to circumnavigate this conclusion. However, all other possible resolutions of the black hole information loss paradox will force us to abandon one of the principles of the underlying theories we are considering, i.e. either features of Quantum Field Theory or General Relativity.

To conclude, however, let me mention that it is a controversial issue whether one should take the black hole information loss paradox as an argument for the non-unitarity of black hole evaporation or as a true paradox. And philosophers have already discussed this issue (Maudlin, 2017a; Unruh and Wald, 1995b, 2017). However, we do want to take a specific stance on this issue since we are more interested in recent developments on black hole paradoxes, which are not crucially dependent on the status of the black hole information loss paradox.

3.2 Page Time Paradox

Page (1993) presents a different paradox which, as we will see, applies long before the full evaporation of the black hole.[10] The paradox will show that, long before the total evaporation, a unitary description of the black hole evaporation conflicts with the Hawking radiation being entirely thermal. Thus, either the black hole evaporates in a non-unitary manner, as Hawking was arguing, or at some point in the evaporation process, the Hawking radiation should start to develop non-thermal features, such as entanglement among its constituents.

To see how the paradox works, we need to review some introductory black hole thermodynamics and its possible statistical mechanics underpinning, which we will do in section §3.2.1. Afterwards, we will review the Page time paradox (§3.2.2).

3.2.1 Black Hole Thermodynamics

Page (1993) considered a black hole to be both a thermodynamic object, as it was already understood by Bekenstein (1973), and a quantum object, i.e. constituted by

[10] See Wallace (2020b) for a detailed philosophical discussion.

some microstates. If this is the case, one can define two types of entropy for black holes. Indeed, a thermodynamic system in a pure state could be described via a microcanonical ensemble and thus possesses a (thermodynamic) entropy S_{MC}, where MC stands for microcanonical:

$$S_{MC}\left(E(t)\right) = \log \dim \mathcal{H}\left[E(t)\right], \tag{3.2}$$

where $\mathcal{H}\left[E(t)\right]$ is the Hilbert space of the system at some time t and energy $E(t)$.[11] On the other hand, given a quantum system, as the black hole would be, one could define the von Neumann entropy[12] of the quantum system

$$S_{VN} = -\mathrm{Tr}\left(\rho \log \rho\right) \tag{3.3}$$

where ρ is the density matrix of that quantum system. One may ask whether these two notions of entropy are somehow connected when speaking of black holes since they are meant to describe the same system from different perspectives. In order to see this, consider the evaporating black hole to be in a pure state before evaporation. Then, by unitarity, at any time t after the evaporation started, the composite system of the black hole and the Hawking radiation should be in a pure state (otherwise, the rules of Quantum Mechanics would be violated). If one considers the radiation and the black hole as two subsystems, then their Von Neumann entropies give the amount of entanglement between them and must be the same since the composite system is assumed to be pure. In particular, S_{VN} increases with the emission of Hawking quanta entangled with the interior. Moreover, we assume that each Hawking quantum is entangled with an interior mode to maintain the black hole plus radiation system in a pure state. Therefore, if black hole evaporation is unitary, S_{MC} bounds S_{VN}, since S_{MC} is proportional to the dimension of the Hilbert space of the black hole (3.2). Indeed, there cannot be more interior modes than the dimensionality of the black hole's Hilbert space. We call this bound the *Page bound*:

$$S_{VN} \leq S_{MC}. \tag{3.4}$$

[11] For a review of quantum statistical mechanics see Mussardo (2010).

[12] Also known as *fine-grained* entropy or *entanglement* entropy. We use these terms interchangeably in what follows.

We have thus seen that the microcanonical entropy bounds the von Neuman entropy in a black hole evaporation process. We will see how this bound generates a paradox at approximately half-time of the black hole evaporation process.

3.2.2 Page Time Paradox

As we have seen in §3.1.1, the black hole cools down through evaporation. It loses energy emitting Hawking quanta in a perfectly mixed state, which means that the microcanonical entropy S_{MC} decreases over time since the dimensionality of the black hole Hilbert space decreases. On the other hand, S_{VN} increases with the number of emitted Hawking quanta since each Hawking quanta is entangled with an interior mode.[13] Therefore, in the evaporation process, there must be a time t_P, called the Page time, at which the bound (3.4) saturates, which means that all interior modes are entangled with a Hawking quantum. The Page time t_P is also when the microcanonical entropy of the radiation becomes larger than the microcanonical entropy of the black hole. The curve, which initially grows with S_{VN} and then decreases with S_{MC}, is called *the Page curve* (see Figure 3.4). Consequently, since after t_P, there cannot be enough interior

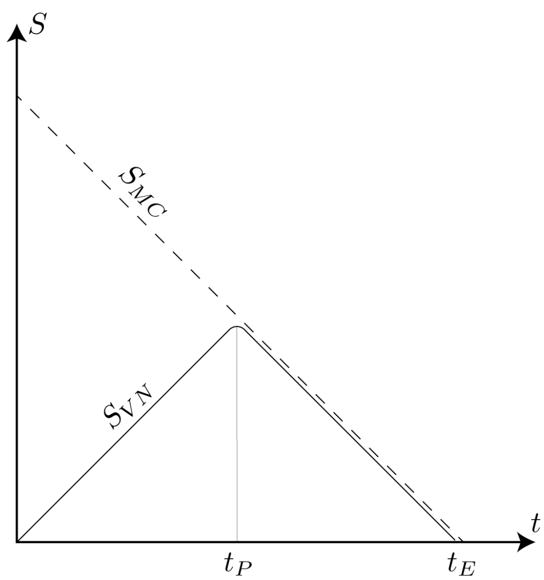


Figure 3.4: The Page curve for an evaporating black hole.

modes to keep the composite system of black hole and radiation in a pure state, one would expect the violation of the bound (3.4) and the non-unitarity of the evaporation

[13]Moreover, also the microcanonical entropy of the Hawking radiation increases with time since its microcanonical entropy is proportional to the number of Hawking quanta.

process. The way to avoid this conclusion is for the late-time Hawking radiation[14] not to be entangled with an interior mode but with something else. The only possibility is that it is entangled with the *early-time radiation*, i.e. the Hawking radiation emitted at times $t < t_P$. In this way, the early radiation purifies the late, keeping the state of the black hole plus radiation system pure. Moreover, the entanglement entropy of the Hawking radiation decreases after t_P, respecting the bound (3.4). The entanglement between early and late radiation implies that Hawking radiation is not perfectly thermal since there are non-trivial correlations among its constituents. Nevertheless, Hawking showed that the radiation *is* perfectly thermal. Therefore, we have a contradiction. This is the nature of PTP: the inconsistency between the prediction of *naive*[15] semiclassical gravity (Hawking's calculation) and black hole statistical mechanics (the entropy bound (3.4)). PTP occurs long before the evaporation time. Indeed, it occurs at the Page time t_p, approximately half of the evaporation time for a Schwarzschild black hole.

A resolution of PTP is offered by the AdS/CFT correspondence, within which one can show that the prediction of black hole statistical mechanics is the correct one.[16] Thus, black hole entropy follows the Page curve, grows with time until t_P and decreases afterwards, and the bound (3.4) is not violated.[17] We will see, however, that such a solution has two caveats: first of all, one should understand why the Hawking result, i.e. the fact that the radiation is perfectly thermal, is wrong after the Page time t_p;[18] second of all, this solution generates a further paradox that we analyse in the next section.

3.3 The AMPS Paradox

While the discussion about PTP has centred around arguments in favour or against the unitarity of black hole physics, said (non) unitarity is not exhaustive of the range of possible black hole paradoxes. In particular, in this section, we describe the firewall

[14] In what follows the words *late* and *early* refer respectively to radiation emitted *after* and *before* the Page time t_P.

[15] Here by *naive* we mean the picture of quantum fields living on the smooth Lorentzian manifold of GR, as used in Hawking (1976).

[16] For a defence of this conclusion see Wallace (2020b). In particular, assuming the correspondence is valid, black hole physics can be shown to be unitary, implying that Hawking radiation cannot be thermal.

[17] As Susskind said "South America wins the war" Susskind (2008), referring to the nationality of Juan Maldacena, the author of Maldacena (1999c), the first article on AdS/CFT.

[18] We will come to this issue later on in the thesis, and we will argue what goes wrong with Hawking radiation after the Page time t_p.

paradox developed in Almheiri et al. (2013) (hereafter, we refer to the authors of this article as AMPS), which threatens the possibility of constructing a consistent theory of the interior. In section §3.3.1, we review a set of reasonable assumptions that Susskind et al. (1993) proposed for a (would be) consistent theory of black holes in Quantum Gravity and section §3.3.2, we will see that these four assumptions conflicts with each other.

3.3.1 Four Principles for Black Hole Physics

Susskind et al. (1993) proposed four principles that a theory that aims to describe the quantum features of black holes should have. These four postulates have been widely accepted among high-energy physicists, at least among physicists in the string theory community. Note that these postulates were originally defined in the context of black hole complementarity. However, they are not strictly tied to this specific proposal. Rather, they are taken to be definitive of what a sensible theory of black hole physics should look like. The four reasonable assumptions that any theory of black holes should satisfy are the following:

Postulate 1 The process of formation and evaporation of a black hole, as viewed by a distant observer, can be described by a unitary S-matrix encoding the evolution from infalling matter to outgoing Hawking-like radiation.

Postulate 2 Outside the stretched horizon[19] of a massive black hole, physics can be described as a good approximation by a set of semiclassical field equations.

Postulate 3 To a distant observer, a black hole appears to be a quantum system with discrete energy levels. The dimension of the subspace of states describing a black hole of mass M is the exponential of the Bekenstein entropy S_{MC}.

Postulate 4 A freely falling observer experiences nothing out of the ordinary when crossing the horizon until the singularity is approached. Another way to say this is that no observer ever detects a violation of the known laws of physics.

The first postulate requires that the theory should describe the time evolution of a black hole in a unitary manner, as any other quantum theory. In particular, this means that the evaporation process of the black hole should be unitary. The second postulate

[19] An horizon at a distance of a Planck length from the true event horizon.

requires that a perturbative quantum gravity theory is enough to describe the physics around the event horizon, which is far away from the singularity. In other words, it requires that the deviations from semiclassical gravity appear only around the singularity of the black hole. The third principle requires that a black hole could be treated as an ordinary thermodynamical object with a statistical mechanics underpinning. The fourth postulate requires that the universality of free fall, required by General Relativity, continues to hold at the event horizon of the black hole. All these principles are pretty reasonable as they stand. However, AMPS showed that these four postulates are inconsistent, in the sense that they generate a paradox. The following section aims to see how the four postulates enter into conflict.

3.3.2 The AMPS Paradox

To see why the four principles for black hole physics proposed by Susskind et al. (1993) are inconsistent, take an evaporating black hole, fix a spatial hypersurface at time $t > t_p$, and consider its Hawking radiation. Divide the black hole plus radiation system into three subsystems: early radiation E, late radiation L and the interior partners to the late radiation B. As we have seen in section 3.2, to have unitary evaporation, as required by Postulate 1, the fact that Hawking radiation is in a pure state implies that the late radiation L and the early radiation E should be entangled. Indeed, a subsystem of a bipartite system in a pure state is maximally entangled with its counterpart. Therefore, since the combined system of early and late radiation is pure (a consequence of Postulate 1), the early and late radiation should be maximally entangled. Moreover, if there is no drama at the horizon (Postulate 4), the late radiation L is fully entangled with the modes behind the horizon B. Indeed, for an infalling observer, the geometry near the horizon can be locally identified with a Rindler horizon via a coordinate transformation:

$$\tau = \frac{t}{4M} \quad , \quad \rho = 2\sqrt{2m(r-2M)} \quad , \quad \tilde{x} = (2M\theta, 2M\phi)\,, \tag{3.5}$$

which turns the Schwarzschild metric into

$$ds^2 \approx -\rho^2 d\tau^2 + d\rho^2 + d\tilde{x}^2, \tag{3.6}$$

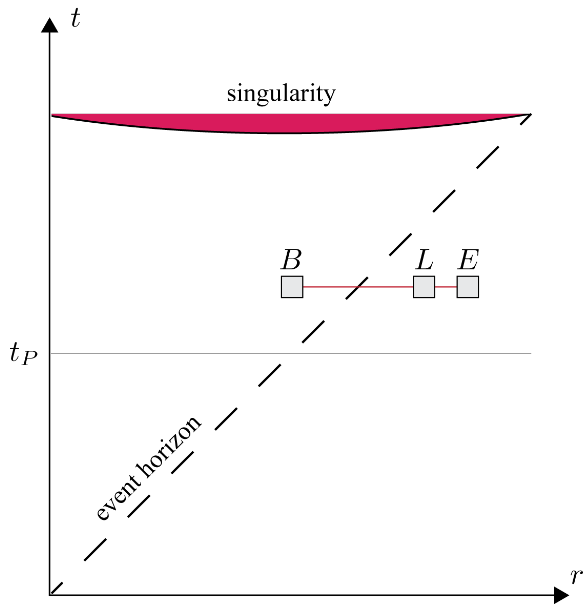


Figure 3.5: A sketch of the basic scenario behind the firewall paradox. B is an interior mode, L is a mode of late Hawking radiation, while E is a mode of early Hawking radiation. The red lines represent relations of maximal entanglement.

the metric corresponding to Rindler spacetime.[20] Fields in the left Rindler wedge are maximally entangled with the right Rindler wedge fields. Therefore a mode L outside the horizon must be maximally entangled with a mode B inside the horizon. Thus the composite system, made of E, L and B, has two features: L is maximally entangled with both E and B.

However, a fundamental property of entangled systems is what in quantum information is called the *monogamy of entanglement*: a quantum system can be maximally entangled with *only one* quantum system at a time. Equivalently, monogamy of entanglement says that there is an upper bound on the independent degrees of freedom with which a given quantum system can be entangled, given by the number of independent degrees of freedom of the quantum system itself. Thus, the four postulates are in contradiction with the monogamy of entanglement. This contradiction is known as the firewall paradox.[21] AMPS proposed that L and B are not entangled but are in a product state. Since L and B being in a product state means that close to the horizon, Rindler space is not a good model for spacetime, assuming that an observer does not see anything out of the ordinary there [Postulate 4] is not reliable. Indeed, if there is no

[20]Note that, here, $d\tilde{x}^2 = d\tilde{x}_1^2 + d\tilde{x}_2^2$, where $x_1 = 2M\theta$ and $x_2 = 2M\phi$.

[21]An equivalent formulation of the paradox is as a violation of strong subadditivity of entanglement entropy, i.e. $S_{AB} + S_{BC} \geq S_B + S_{ABC}$.

entanglement between L and B, B is not entangled with anything. On the other hand, L is still entangled with E, which purifies it. Therefore the state of B is characterised by its reduced density matrix, which is thermal.[22] The thermal nature of B leads to a high concentration of energy at the horizon, which can be suggestively described as a wall of fire, hence the notion of the firewall.

The existence of firewalls implies that a non-trivial theory of the interior is impossible since the interior would always be characterised by the same thermal state, i.e. the firewall. In other words, no interesting theory of the black hole interior would be possible in the presence of firewalls. In the following chapters of this thesis, we will analyse the space of possible solutions to this paradox. We will see that there are more conservative solutions in which we do not have to give up on a consistent theory of the black hole interior.

3.4 The AMPSS Paradox

This section aims to present the AMPSS paradox, which can be understood as a generalisation of the AMPS paradox. The AMPS paradox, analysed in the previous section, showed an inconsistency between the principles proposed by Susskind et al. (1993) for a consistent theory of quantum black holes and the monogamy of the entanglement. The AMPSS paradox, instead, concerns a possible obstruction to the program of holographic interior reconstruction, i.e. the idea that the interior of a black hole can be encoded in a lower dimensional boundary CFT. In this regard, as we were also saying at the beginning of this chapter, this last paradox can be better understood in the context of holography, particularly in the AdS/CFT correspondence. The application of the AdS/CFT correspondence to black hole physics is the following: since a gravitational theory in AdS should be equivalent to a non-gravitation theory in the boundary of AdS, one might think to describe a black hole in AdS, also at a quantum level, by a quantum non-gravitational theory living in the boundary of AdS. In particular, one could *in principle* be able to describe the interior of a black hole, also during its evaporation process, by a non-gravitational conformal field theory. The research program that aims to describe the physics of the interior of black holes using AdS/CFT techniques is called

[22] The thermal nature of the state of B is because, since we want the entangled and non-entangled descriptions to be locally indistinguishable, to describe the state of B, we use the reduced density matrix of the maximally entangled state of the composite system of B and L. It is a well-known fact that the reduced density matrix of a subsystem of a maximally entangled system is completely thermal.

holographic interior reconstruction.

To understand the AMPSS paradox, we will first introduce, in section §3.4.1, the basics of the program of holographic interior reconstruction (Harlow, 2018) and, in particular, we give a brief introduction to entanglement wedge reconstruction, which is the primary approach for describing the interior in the boundary CFT. In §3.4.2, we look at entanglement wedge reconstruction for evaporating black holes, following Penington (2020). Finally, in section §3.4.3, we present the AMPSS paradox.

3.4.1 Entanglement Wedge Reconstruction[23]

In this section, our primary goal is to introduce the basics of entanglement wedge reconstruction, which gives a prescription within the AdS/CFT correspondence for which a portion of the bulk is encoded in a given subregion of the boundary.[24] The first notion that we need to define is that of *quantum extremal surface* (see Figure 3.6):

(QES) A **quantum extremal surface** χ is defined as a surface satisfying two conditions:

(i) *Homology Constraint:* given a boundary region B, a surface χ satisfies the homology constraint if, for C a space-like hypersurface, $\chi \cup B = \partial C$, i.e. the union of χ with a boundary region B is the boundary of some space-like region C. C is called *homology hypersurface.*

(ii) *Extremize the Generalised Entropy:* the surface χ should be a surface which extremizes the generalised entropy

$$S_{\text{gen}}(\chi) = \text{ext}\left[\frac{A(\chi)}{4G_N} + S_{\text{bulk}}(\chi)\right] , \tag{3.7}$$

where $S_{\text{bulk}}(\chi)$ is the von Neumann entropy[25] of the bulk fields contained in $\chi \cup B$ and $A(\chi)$ is the area of the hypersurface χ.

While $S_{\text{bulk}}(\chi)$ in general diverges (and should be renormalized), $S_{\text{gen}}(\chi)$ remains finite since $A(\chi)$ has a contribution that directly cancels the divergence of $S_{\text{bulk}}(\chi)$.

[23] For background material on entanglement wedge reconstruction see Harlow (2018). For philosophical discussion on the Ryu-Takayanagi formula and related issues, see Jaksland (2018), and Bain (2020a).

[24] Note that, from now on, when speaking of the boundary CFT, we use the terms *subsystem* and *subregion* interchangeably.

[25] If the total bulk state is pure, then this von Neumann entropy is the entanglement entropy between what is inside and what is outside of the entanglement wedge.

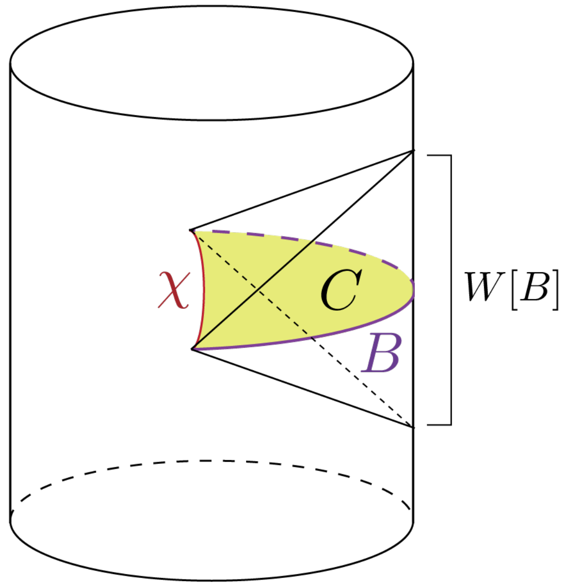


Figure 3.6: An example of entanglement wedge. Here $W[B]$ is the entanglement wedge of the boundary region B (purple), χ (red) is the **QES** which minimises S_{gen}, and C (yellow) is the homology hypersurface.

The Ryu-Takayanagi prescription states that the entanglement entropy S_{EE} of some spatial subregion B of the CFT is the generalised entropy S_{gen} computed on the **(QES)** which minimises it. We call this **(QES)** the *quantum Ryu-Takayanagi* or *HRT*[26] surface. In particular, for a generic quantum system in a pure state, if one divides the system into two subsystems A and B, the entanglement entropies of A and B are the same, i.e. $S_{\text{EE}}(A) = S_{\text{EE}}(B)$. Given the RT prescription for computing the entanglement entropy holographically, the two entanglement entropies are the same means that both regions have the same HRT surface.

We can now give a precise definition of the *entanglement wedge* (see Figure 3.6):

(EW) Let B be a boundary spatial subregion of an asymptotically-AdS spacetime. The **entanglement wedge** of B, which we denote by $W[B]$, is the bulk domain of dependence $D[C]$[27] of the homology hypersurface C delimited by the HRT surface χ.

[26]From the initials of Hubeny et al. (2007), who first introduced it.

[27]The domain of dependence $D[C]$ is the set of points with the property that any causal curve passing through one of these points must also intersect C.

Then, **(EW)** leads to the following *entanglement wedge reconstruction* conjecture:[28]

(EWR) **Entanglement wedge reconstruction** says that all physical quantities in $W[B]$, i.e. the entanglement wedge of a spatial subregion B, are represented in the CFT by operators in B.

In other words, the entanglement wedge of some spatial boundary region B is the portion of the bulk encoded in B. Any bulk operator in $W[B]$ can be reconstructed by operators in the boundary region B, and, in general, this reconstruction procedure takes the form of an error-correcting map.[29] Moreover, the entanglement wedges of boundary regions with the same HRT surface, i.e. regions whose composite state is pure, are complementary.

Note that entanglement wedge reconstruction allows us to move beyond the causal structure of general relativity. Indeed, for relativistic bulk geometries, besides $W[B]$, one can also define the causal wedge $C[B]$ of some boundary region B. $C[B]$ is the intersection of the bulk future and the bulk past of $D[B]$, i.e. the domain of dependence of B. While the entanglement wedge $W[B]$ and the causal wedge $C[B]$ coincide for a ball-shaped boundary region in empty AdS, they are generally different. In particular, it is always the case that

$$C[B] \subseteq W[B] \ . \tag{3.8}$$

Indeed, it is never the case that the interior of a black hole is in the causal wedge of a boundary region B since the interior and the exterior of a black hole are causally disconnected. On the other hand, as we will see in the next section, the black hole interior is included in the *entanglement wedge* of a boundary region B.

[28]To be precise, assuming the AdS/CFT correspondence, **(EWR)** is a theorem (Jafferis et al., 2016; Dong et al., 2016; Harlow, 2017). However, beyond AdS/CFT, it remains a conjecture.

[29]Error-correction in AdS/CFT has been introduced by (Almheiri et al., 2015). For philosophical discussion of the role of error correction in AdS/CFT, see Bain (2020b).

3.4.2 The EW Structure of an Evaporating Black Hole[30]

In this section, we are now going to study **(EWR)** for a large[31] evaporating black hole in AdS formed from gravitational collapse. This case is treated in great detail by Penington (2020), whom we follow in this section. Large black holes in AdS do not usually evaporate since Hawking radiation reaches the boundary and reflects into the black hole, quickly reaching thermal equilibrium; the number of outgoing and ingoing Hawking quanta is the same. This fact is due to the specific boundary conditions of AdS (usually called *reflecting boundary conditions*, see Figure 3.7a). Therefore, large black holes in AdS do not spontaneously evaporate. However, black holes can evaporate by changing the boundary conditions of AdS and allowing some Hawking radiation to escape the boundary. We call these new boundary conditions *absorbing boundary conditions* (see Figure 3.7b), and we call $\mathcal{H}_{\text{rad}}$[32] the extra system which stores the escaping Hawking radiation.

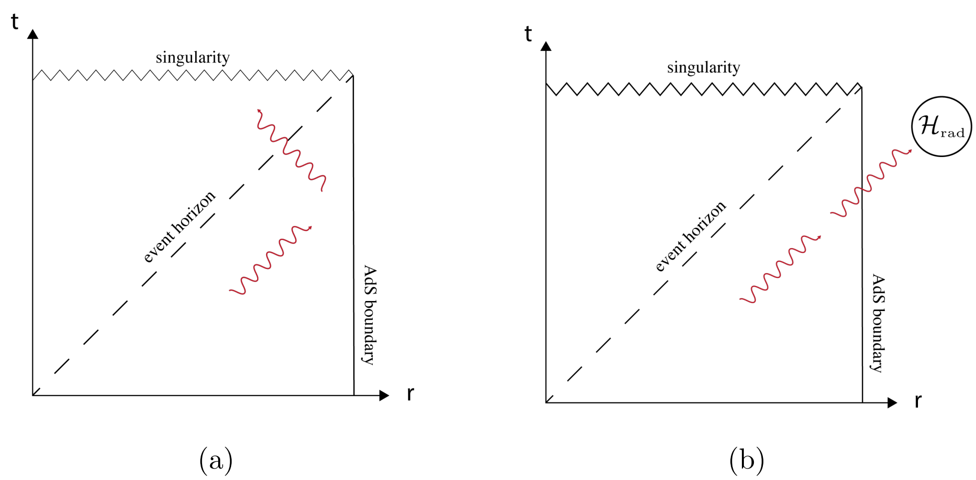


Figure 3.7: (3.7a) is the Penrose diagram for a black hole with reflecting boundary conditions. Observe that the radiation (wiggly red lines) bounces on the boundary and returns to the black hole. (3.7b) is the Penrose diagram for a black hole with absorbing boundary conditions. Observe that some of the radiation (wiggly red lines) escapes from the boundary and is holographically encoded in $\mathcal{H}_{rad}$.

From a boundary perspective in AdS/CFT, this procedure is the same as (generically)

[30]This subsection is rather technical. The motivation for this section is the lack of an appropriate technical introduction to entanglement wedge reconstruction for evaporating black holes in the philosophical literature. Hose who are not interested in the details of these constructions should skip directly to the conclusion of the section. Keep in mind that the main takeaway here is that there are operators in the holographic dual $\mathcal{H}_{\text{rad}}$ of the Hawking radiation representing the black hole interior.

[31]Here large is defined with respect to the radius of curvature of AdS spacetime.

[32]With a slight abuse of notation we refer to quantum systems and their Hilbert space interchangeably.

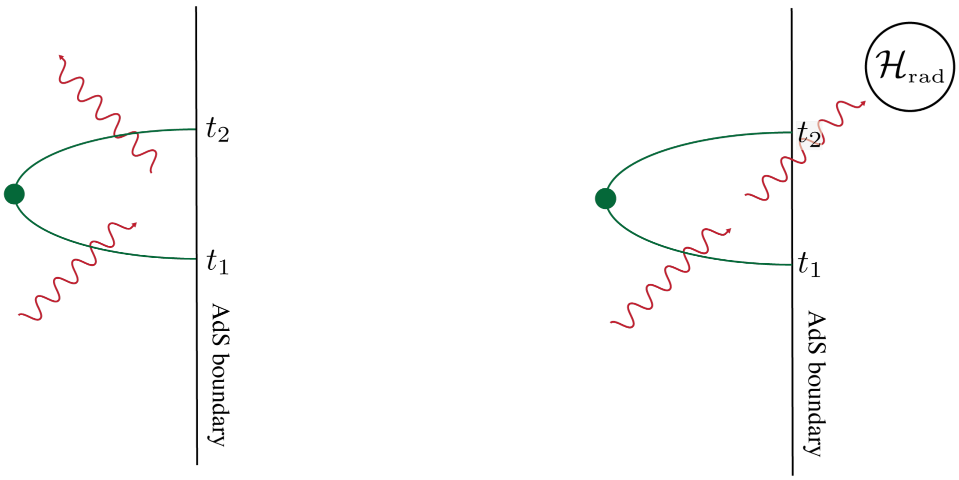
coupling the boundary CFT $\mathcal{H}_{\text{CFT}}$ to the auxiliary system $\mathcal{H}_{\text{rad}}$.[33] We also assume that $\mathcal{H}_{\text{rad}}$ is itself a large holographic system with a (not further specified) gravitational dual. A natural question within **(EWR)**, crucial in finding a solution to the AMPSS paradox, is which part of the bulk is encoded in $\mathcal{H}_{\text{CFT}}$ and which part of the bulk is encoded in $\mathcal{H}_{\text{rad}}$. Let us discuss this point in detail. Since we have two boundary CFTs, i.e. $\mathcal{H}_{\text{CFT}}$ and $\mathcal{H}_{\text{rad}}$, we have two entanglement wedges, one for $\mathcal{H}_{\text{CFT}}$ and one for $\mathcal{H}_{\text{rad}}$. However, since $\mathcal{H}_{\text{rad}}$ and $\mathcal{H}_{\text{CFT}}$ encode all the bulk physics (as their union is the entire boundary), and the overall bulk state is pure, $\mathcal{H}_{\text{rad}}$ and $\mathcal{H}_{\text{CFT}}$ have the same HRT surface, and their entanglement wedges are complementary, in accord with our discussion in §4.2. Therefore, a generic bulk mode is either in the entanglement wedge of $\mathcal{H}_{\text{rad}}$ or the entanglement wedge of $\mathcal{H}_{\text{CFT}}$.

In the context of absorbing boundary conditions, which bulk mode is in which the entanglement wedge depends on the boundary time. Indeed, consider two extremal surfaces anchored respectively at boundary times t_1 and t_2. With reflecting boundary conditions, the same degrees of freedom are always contained in the same surfaces, since outgoing modes in an extremal surface anchored at time t_1 are just ingoing modes in a surface anchored at time t_2 (see Figure 3.8a). However, with absorbing boundary conditions, it can happen that outgoing modes at time t_1 escape into $\mathcal{H}_{rad}$ and do not appear (as ingoing modes) in the surface anchored at t_2 (see Figure 3.8b). Hence bulk entropy, **(QES)**, and the HRT surface are *time-dependent* notions when dealing with absorbing boundary conditions. In other words, it can happen that a **(QES)** for the entire boundary at time t_1 is not anymore a **(QES)** for the entire boundary at time t_2.

Let us now look at an evaporating black hole, holographically dual to $\mathcal{H}_{\text{rad}} \otimes \mathcal{H}_{\text{CFT}}$. Here we have two different **(QES)**: the empty surface $\varnothing$, i.e. a surface such that we associate to $\mathcal{H}_{\text{CFT}}$ the entirety of the bulk spacetime, and the surface $\bar{\chi}$, which lies just inside the event horizon.[34] Note that, since the overall state of the black hole is pure, $\mathcal{H}_{\text{rad}}$ and $\mathcal{H}_{\text{CFT}}$ must have the same **(QES)** as their HRT surface, as seen in §4.2. To find which is the HRT surface, we need to compute the generalised entropy S_{gen} defined by (3.7), which is given by two different contributions: an area term A and a bulk entropy term S_{bulk}. In particular:

[33] That Hawking radiation only goes from $\mathcal{H}_{\text{CFT}}$ to $\mathcal{H}_{\text{rad}}$ and not vice-versa follows from a generic boundary state $|\psi\rangle$ in $\mathcal{H}_{\text{CFT}}$ obeying a Markovian master equation (Preskill, 2015).

[34] For a proof of their extremality see Penington (2020).



(a) (b)

Figure 3.8: (3.8a) is a **(QES)** in AdS spacetime with reflecting boundary conditions, where the hypersurface is not time dependent. (3.8b) is a **(QES)** in AdS spacetime with absorbing boundary conditions, where the hypersurface is time dependent.

$S_{\text{gen}}(\varnothing) = S_{\text{rad}}$ since the area term $A(\varnothing)$ is zero while S_{bulk} is given by S_{rad}. Indeed, the area of an empty surface is trivially zero. On the other hand since the empty surface associates to $\mathcal{H}_{\text{rad}}$ only the escaped Hawking radiation whose entropy is S_{rad}, and the state of the composite system of $\mathcal{H}_{\text{rad}}$ and $\mathcal{H}_{\text{CFT}}$ is pure, then S_{rad} must be the von Neumann entropy S_{bulk} of $\mathcal{H}_{\text{CFT}}$ and $\mathcal{H}_{\text{rad}}$. Thus, $S_{\text{gen}}(\varnothing) = S_{\text{rad}}$.

$S_{\text{gen}}(\bar{\chi}) \approx \frac{A_{\text{hor}}}{4G_N}$ since the area term approximately corresponds to the black hole area while S_{bulk} is zero. Indeed, since $\bar{\chi}$ sits approximately at the horizon, its area is approximately A_{hor}. On the other hand, while the only contribution to the von Neumann entropy S_{bulk} comes from the Hawking radiation, the escaped Hawking quanta and the interior modes with which they are entangled are both in the same entanglement wedge (that of $\mathcal{H}_{\text{rad}}$, as defined by $\bar{\chi}$), thus making the overall von Neumann entropy of $\mathcal{H}_{\text{rad}}$ zero. Again, however, since the composite state is pure, the von Neumann entropy S_{bulk} is zero also for $\mathcal{H}_{\text{CFT}}$. Thus, $S_{\text{gen}}(\bar{\chi}) \approx \frac{A_{\text{hor}}}{4G_N}$.

Before the Page time, the following equation holds

$$\frac{A_{\text{hor}}}{4G_N} > S_{\text{rad}} \ , \tag{3.9}$$

equivalently, the microcanonical entropy of the black hole, given by the area, is greater than the entropy of the Hawking radiation. Indeed, the Page time is the time at which equation (3.9) becomes an equality. Thus, since the horizon area is greater than S_{rad}, $S_{\text{gen}}(\varnothing)$ is smaller than $S_{\text{gen}}(\bar{\chi})$, and the HRT surface is the empty surface since it minimises S_{gen}. S_{rad} here gives Hawking (1976)'s semiclassical thermal entropy, which is the right entropy before the Page time but not after.[35] Thus, since $\varnothing$ does not associate any bulk region with $\mathcal{H}_{\text{rad}}$, the interior is still in the entanglement wedge of the CFT, not in that of $\mathcal{H}_{\text{rad}}$. Before the Page time, we could not reconstruct the interior from the radiation, and the information has not escaped.

After the Page time, the converse of (3.9) holds, and the empty surface, even though it is still extremal, is not the HRT surface anymore. Indeed, since S_{rad} is greater than A_{hor}, $S_{\text{gen}}(\bar{\chi})$ is smaller than $S_{\text{gen}}(\varnothing)$. Thus, it is the surface $\bar{\chi}$ which minimises S_{gen} and is the new HRT surface.

Hence, in the black hole evaporation process, we have a phase transition between HRT surfaces: before the Page time, the empty surface is the HRT surface while, after the Page time, the new HRT surface is the non-empty surface $\bar{\chi}$, which lies just inside the event horizon. A consequence of this process is that the following formula gives the entanglement entropy between the radiation and the CFT:[36]

$$S_{EE} = \min\left(S_{rad}, \frac{A_{hor}}{4G_N}\right) . \tag{3.10}$$

The phase transition between $\varnothing$ and $\bar{\chi}$ implies that before and after the Page time, we must consider two different entanglement wedges for $\mathcal{H}_{\text{rad}}$ and $\mathcal{H}_{\text{CFT}}$, since the entanglement wedge, as defined by **(EW)**, is delimited by the HRT surface. For information to escape from the interior, by **(EWR)**, it should be in the entanglement wedge of $\mathcal{H}_{\text{rad}}$ because $\mathcal{H}_{\text{rad}}$ is the only subsystem accessible to an observer outside the black hole as it is the subsystem representing the Hawking radiation. Since, before the Page time, $\varnothing$ is the HRT surface, the interior is in the entanglement wedge of the CFT. Therefore, no information can escape outside since the interior is not in the entanglement wedge of $\mathcal{H}_{\text{rad}}$. However, after the Page time, since $\bar{\chi}$ is the new HRT surface, part of the interior is in the entanglement wedge of $\mathcal{H}_{\text{rad}}$. This phase transition allows the information of the interior to escape since by **(EWR)** we can reconstruct it starting from $\mathcal{H}_{\text{rad}}$ and thus from the Hawking radiation available to an observer outside the black hole (see Figure 3.9). The interior region contained in $W[\mathcal{H}_{\text{rad}}]$, which is spacelike separated from

[35] For a defence of this claim, see Wallace (2020a).

[36] From this formula, we can derive the Page curve (Page, 1993). Hence, Hawking (1976)'s semiclassical calculation of S_{EE} is only accurate before the Page time, not after.

$\mathcal{H}_{\rm rad}$, is called in the literature *(quantum) island* (Almheiri et al., 2020b).[37]

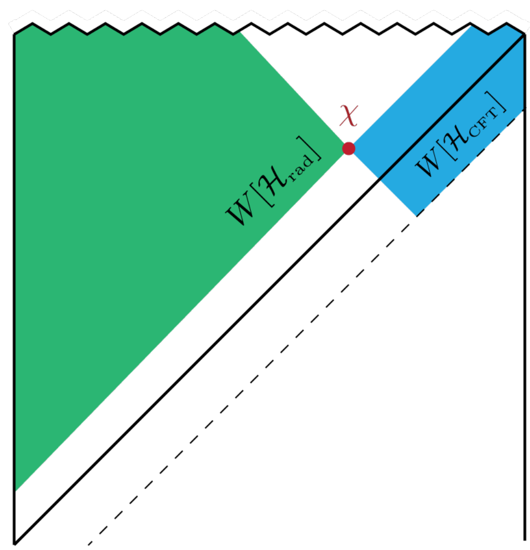


Figure 3.9: The Penrose diagram of an evaporating black hole after the Page time. The green region is the entanglement wedge of $\mathcal{H}_{\rm rad}$, while the blue region is the entanglement wedge of $\mathcal{H}_{\rm CFT}$. The HRT surface χ is in red. It is clear from the picture that (most of) the interior of the black hole is contained in the entanglement wedge of $\mathcal{H}_{\rm rad}$.

The upshot of this section is that there should be operators in $\mathcal{H}_{\rm rad}$ that represent interior bulk quantities. In other words, via **(EWR)**, $\mathcal{H}_{\rm rad}$ is capable of peering through the black hole horizon and seeing the interior. However, in the next section, we will see an argument that such operators cannot exist.

3.4.3 The AMPSS Paradox

The AMPSS paradox, also known as the creation operator problem, much like holographic interior reconstruction, is most easily stated in the AdS/CFT correspondence context. To start, take a large black hole in AdS spacetime. The paradox consists of an argument that, in the boundary CFT, it is inconsistent to have an operator $b^\dagger$, which

[37] While careful discussion of this point goes beyond the scope of this article, let us mention that Geng et al. (2021) suggest that concrete models of interior reconstruction in $d > 2$ tend to imply that the gravity in bulk is not GR but a theory of Massive Gravity. It is, however, unclear whether this is a mere technical limitation of the models studied thus far or a deep fact about holographic interior reconstruction.

acts as a creation operator for modes in the interior of the black hole. Thus, if this were the case, interior reconstruction would be impossible since we could not express interior quantities in the boundary CFT, contra the conclusions of section §3.4.2. Almheiri et al. (2013) state the paradox for general, large black holes in AdS and, in this section, we present it as such. However, in the following parts of this thesis, we are going to specialise our discussion of the paradox in the context of evaporating black holes most relevant to interior reconstruction, where the goal is to reconstruct the interior from $\mathcal{H}_{\text{rad}}$. For discussion of the same constructions that we study here in the context of more general types of black holes, which go beyond the scope of the present article, see Almheiri (2018), and Papadodimas and Raju (2016).

Observe first of all that $b^\dagger$ must lower the energy in the CFT, which implies that $b^\dagger$ satisfies the following equation:

$$[H, b^\dagger] = -\omega b^\dagger \ , \tag{3.11}$$

where H is the CFT's Hamiltonian, and ω is an energy eigenvalue of H. To get an intuition for this claim, note that in the case of an evaporating black hole, $b^\dagger$ must be involved in the process of the emission of Hawking radiation, as it creates the interior modes entangled with the exterior radiation quanta. Since the Hawking radiation reduces the black hole's energy and increases the exterior's energy, i.e. the black hole evaporates, $b^\dagger$ must lower the energy in the boundary CFT in which it lives.

However, $b^\dagger$ is also supposed to be a standard creation operator in the boundary CFT. In QFT, creation operators must have a left inverse[38] b. The existence of a left inverse is equivalent to there being an operator b satisfying the following constraint:

$$\left(\frac{1}{1+bb^\dagger}b\right)b^\dagger = 1 \ . \tag{3.12}$$

However, (3.11) and (3.12) are incompatible. Indeed, since $b^\dagger$ lowers the energy of the CFT, it is a map from a subspace of CFT states of energy E to a subspace of states of energy $E-\omega$. For ω big enough, the difference between the two subspaces is $O(1)$, which implies that $b^\dagger$ cannot be invertible since it is a many-to-one map, not one-to-one. However, $b^\dagger$ must be invertible by (3.12); otherwise, it cannot have a left inverse. Thus we have a contradiction that seems to make the existence of $b^\dagger$ in the CFT impossible.

[38]The left inverse of an operator $\mathcal{O}^\dagger$ is an operator $\mathcal{O}$ such that $\mathcal{O}\mathcal{O}^\dagger = \mathbf{I}$, the identity operator.

To get a physical intuition for the incompatibility between (3.11) and (3.12), note that, since $b^\dagger$ lowers the energy of the CFT, at some point, it will reach a subspace of states with minimal energy, i.e. the vacuum. Applying again $b^\dagger$, we should then annihilate the vacuum, i.e. $b^\dagger |0\rangle = 0$, where $|0\rangle$ is the vacuum state. However, this cannot be the case since $b^\dagger$ is a creation operator and thus cannot annihilate the vacuum, i.e. $b^\dagger |0\rangle \neq 0$.[39] Thus, we have a contradiction between (3.11) and (3.12).

Almheiri et al. (2013) suggest that we respond to this situation by abandoning our initial purpose of describing the interior within the boundary CFT. Instead, we should assume a firewall at the black hole horizon, which makes the interior *non-existent*, explaining why it cannot be studied from the CFT perspective. In this way, one would solve the AMPSS paradox by simply postulating that there should not be any boundary operator $b^\dagger$ because the interior, and *a fortiori* any operator localised there, cannot be reconstructed within the CFT. However intriguing, this solution has two big drawbacks: it makes the interior of the black hole impossible to study and the information in the interior impossible to retrieve since it lies beyond the firewall. Since these were the two main goals from which the whole project of holographic interior reconstruction started, we take that the firewall solution should, at best, be seen as an *extrema ratio* response to AMPSS. However, we will see, in the following chapters, that it is possible to overcome the AMPSS paradox while still describing the interior in the boundary CFT and thus retain the primary hope behind the program of holographic interior reconstruction.

[39] This statement follows from the fact that $b^\dagger$, acting on a generic state, creates modes, which is incompatible with annihilating the vacuum.

Chapter 4

A Philosophical Tool: Causal Structures

This chapter aims to develop a philosophical formalism that we can use to analyse the conceptual foundation of the various black hole paradoxes we analysed in the previous chapter. Indeed, as it was already clear when reviewing paradoxes in modern physics in chapter §3, once we explicitly find a paradoxical situation, we need to understand which principles enter into conflict. And we can do so only if we find a way to formulate the physical principles explicitly. From the point of view presented in §4, and from the formal analysis of black hole paradoxes one can find in a physics paper, it is not clear which are all the principles responsible for the physical situation under study and, as a consequence, it is not clear where the conflict between these principle stands. As argued in §3, I believe this is a task for philosophers of physics and can contribute to cutting-edge research because it can help to elucidate not only where the conflict stands but also which are the possible strategies to overcome the conflict. In this chapter, we will define the philosophical tool we need to highlight the conflict in the various black hole paradoxes, what we call *causal structure*. In the next chapter, we will see how the philosophical tool we introduce helps to highlight the various conceptual strategies that can be used to solve black hole paradoxes and how these strategies match with proposed solutions in contemporary theoretical physics literature.

To understand the potential conflict between General Relativity and Quantum Mechanics (or Quantum Field Theory), it is important to first consider the differences between the two theories. These two theories are very different in many respect. Firstly, General Relativity is a classical theory that describes gravity as the curvature of spacetime caused by the presence of matter and energy. On the other hand, Quantum Mechanics is a quantum theory that describes the behaviour of matter and energy at a subatomic level. The difference between these two theories is that General Relativity deals with macroscopic objects and the large-scale structure of the universe, while Quantum Mechanics deals with the microcosm of subatomic particles. Secondly, General Relativity is a deterministic theory, which means that the state of a system at a given time is uniquely determined by its state at an earlier time and the laws of physics. On the other hand, Quantum Mechanics is a probabilistic theory, which means that the state of a system can only be described in terms of probabilities. This is because the behaviour of subatomic particles is inherently uncertain and cannot be predicted with complete accuracy. The fact that General Relativity is a classical and deterministic theory, while Quantum Mechanics is a quantum and probabilistic theory, makes it difficult to reconcile the two theories in the context of black holes, which are objects that exist at the boundary between the macroscopic and the subatomic realms, and between the deterministic and the probabilistic. Resolving the black hole paradoxes will require a deep understanding of the fundamental principles of both General Relativity and Quantum Mechanics and a novel framework that can reconcile their differences.

However, the differences we are interested in, and which are going to be crucial in analysing black hole paradoxes, are two:

Distinctness between systems, i.e. given two systems, when we can say they are distinct. In General Relativity, two systems are considered distinct if they have support in non-overlapping spacetime regions. This means that the two systems do not share any common points in spacetime and hence can be distinguished from each other. This distinction is based on the geometric properties of spacetime and the presence of matter and energy in different regions.[1] In Quantum Mechanics, systems are typically identified through their algebra of observables, which are the set of measurable quantities associated with a system. Two systems are considered distinct if their algebra of observables commutes, meaning that any measurement on one system is independent of any measurement on the other. This is because mutual commutativity guarantees that the two systems have no degrees of freedom in common. Identifying distinctness with mutual commutativity is relatively

[1]Note that we are not taking any particular stance toward General Relativity ontology here. The word *support* is intended in the mathematical sense of the term.

standard in quantum theory.[2] Conversely, if the algebras of observables do not commute, the two systems have some degrees of freedom in common, making it difficult to distinguish between them. This is because observables with non-zero commutators are not independent, meaning that determining the value of one observable entails fixing the system's state for both observables. This implies that the two systems are not distinct and cannot be treated independently. An example of this is a one-dimensional system with only position and momentum degrees of freedom. In quantum mechanics, the state space of such a system is given by $L^2(\mathbb{R})$, and specifying the system's position fixes its wave function and, thus, in particular, its momentum state. This is in contrast to classical mechanics, where the specification of the system's position is completely independent of its momentum. This illustrates the fundamental differences in how systems are distinguished in General Relativity and Quantum Mechanics and highlights the importance of understanding the underlying principles of each theory when trying to distinguish systems.

Connections between systems, i.e. given two distinct systems, when we can say they are connected. In General Relativity, the connection between two systems is determined by the lightcone of one system in relation to the other. Specifically, two systems are considered connected if the spatiotemporal distance between them is either time-like or null, meaning they can be connected by a causal path. On the other hand, if the distance between the two systems is space-like, they are not connectable. In Quantum Mechanics, the connection between two systems is more subtle because it heavily depends on the ontology of quantum mechanics one chooses. One way to understand connections between systems in quantum mechanics could be through entanglement, where two distinct systems can be connected through a shared quantum state. However, the precise definition of connection in quantum mechanics is still a topic of ongoing research and debate. We are going to discuss this issue later in this chapter. For the time being, we consider entanglement between two distinct systems as a type of connection between two distinct systems in Quantum Mechanics.

These differences between General Relativity and Quantum Mechanics will be crucial in analysing black hole paradoxes and will allow us to elucidate the conceptual structure of

[2] Though see Earman and Valente (2014) for some subtleties about this claim.

the AMPS and AMPSS paradoxes. In particular, we are going to introduce a new notion that we call *causal structure*, that will rely on the idea of connectedness between two distinct systems in a particular theory (§4.1). With this notion in mind, we are going to analyse the *Spacetime causal structure* (§4.2)and the *entanglement causal structure* (§4.3). Moreover, we will see that these two causal structures are compatible far from black holes. A the end of this chapter, we are going to introduce a new causal structure that we are going to call *generalised causal structure* or *wormhole causal structure* (§4.3.3),[3] which is going to be useful elucidating one of the possible resolutions of the black hole paradoxes.

4.1 Causal Structures

Keeping in mind the difference we are interested in between General Relativity and Quantum Mechanics, we would like to have a tool that can keep track of and highlight these differences. We want a definition that captures the connections between two distinct entities of a specific theory. We, thus, define *causal structure* in the following way:

(CS) **Causal Structure**: given a theory T, we say that the causal structure according to the theory T is provided by a set of spacetime regions/objects (with their physical state) and a relation R which determines if two objects/regions of spacetime can or cannot be causally related.

In this definition, we intend to identify causally robust counterfactual connections. By robust counterfactual connections, we mean connections which give rise to appropriate counterfactuals, i.e. expressions of the form "If A had not occurred, C would not have occurred".[4] We use robust counterfactual connections to exclude spurious correlations with no physical significance. Spurious correlations are those that do not give rise to reliable counterfactuals. For example, assume that the price of bread in England is roughly correlated with the water levels in Venice. Both tend to rise. However, it is not the case that if the price of bread in England were to fall, then water levels in Venice would likewise decrease. Hence, spurious correlations do not give rise to

[3] Why we are dubbing this particular causal structure the wormhole causal structure will become clear later on in this thesis.

[4] See Lewis (2013) for a classic study of counterfactuals and Starr (2021) for a review of the literature.

reliable counterfactuals. Moreover, counterfactuals can be used to construct a theory of causality, based, very roughly, on the idea that "A causes B" if and only if "if A had not occurred, B would not have occurred" (Lewis, 1974; Menzies and Beebee, 2020). We identify causality with robust counterfactual connections only for ease of exposition to avoid cumbersome phrases such as *robust counterfactual connections*. The reader who dislikes this identification or has more stringent requirements on causation is free to substitute causality talk with robust counterfactual connection talk. The crucial point for us is that, in the definition **(CS)**, causality is intended to be robust counterfactual connections. Otherwise, we could not define a **(CS)** of the entanglement relation, which is crucial for our reconstruction of black hole paradoxes. Thus, the reader who prefers a stronger notion of causality is free to substitute for our talk of causality talk of robust counterfactual connections.

With this definition in hand, we consider two different causal structures based on two different triples: a theory T, a set of spacetime regions/objects Ω, and a relation R. In section §4.2, we define the *Spacetime Causal Structure*, which is the Causal Structure of General Relativity. In section §4.3, we describe the *Entanglement Causal Structure*, which is the Causal Structure of a Quantum Theory.

4.2 The Spacetime Causal Structure

As we mentioned in the introduction of this chapter, **(CS)** could depend on the ontology one considers for the particular theory T considered in the definition of **(CS)**. This is because, in the definition, we explicitly refer to the set of spacetime regions/objects (with their physical state), which could depend on one's preferred ontology of that particular theory T. In the case under consideration, it is not obvious how GR should be interpreted. Here, however, for exposition, we speak explicitly of spacetime points. Nonetheless, it should be possible to carry over this discussion for other approaches to GR's ontology, at least insofar as a notion of object is admitted in the ontology. For the time being, let us consider that the set of objects constitutive of General Relativity is spacetime points. We do this mainly for ease of exposition, to have something explicit to consider, such as spacetime points. Nonetheless, it should be possible to carry over this discussion for other approaches to GR's ontology, at least insofar as a notion of object is admitted in the ontology.

Take as our physical theory T GR, with its set of objects Ω being the spacetime points, and identify the relation R in **(CS)** with the relation R_{LC} of *being connectable by a causal curve*, obtaining between two points of spacetime p and q. We understand the relation R_{LC} as obtaining if and only if there is a causal curve between the two points of spacetime p and q. By causal curve, we mean a timelike or null curve (that is, a curve which is always within the light-cone, border included, or, equivalently, a curve representing the worldline of an object whose speed never exceeds the speed of light). We call this the *causal structure of (relativistic) spacetime*.[5] Thus, causal curves describe the causal structure of spacetime in the sense that they determine which points can be in causal contact. Note that with this relation, one automatically encodes the locality properties of relativity theory since it follows that there cannot be spacelike separated points which are in causal contact.

4.3 The Entanglement Causal Structure

Defining a causal structure **(CS)** of entanglement would require two steps: first of all, one needs to identify the set of objects of a given quantum theory; second of all, one should be able to interpret entanglement as a relation, thus having a concrete relation R to use in the definition **(CS)**. Before moving on to the actual definition of The Entanglement Causal Structure and its consequences (§4.3.2), we would like to set the stage for it by giving a possible interpretation of entanglement as a relation (§4.3.1).

4.3.1 Entanglement as a Relation[6]

A fundamental task for making sense of Quantum Mechanics from a metaphysical point of view is providing an interpretation of entanglement. Indeed, the experimentally confirmed correlations – which entanglement states give rise to – cannot be explained in terms of the intrinsic properties of the quantum systems involved. How to explain these correlations is considered a metaphysical conundrum for those who seek a local realist ontology of QM.

Cinti et al. (2022) follows entanglement realists in articulating a positive metaphysics of entanglement and explores one alternative to wave-function realism according to

[5] In what follows, we omit the relativistic qualification as we always talk about relativistic spacetimes.

[6] The content of this section is a short and rough summary of Cinti et al. (2022). For a detailed analysis of why one could interpret entanglement as an (external) relation, I recommend the reading of Cinti et al. (2022).

which entanglement relations are physical relations which "do not supervene on the non-relational physical properties" (Teller, 1986, p.73) of the entangled systems. Relational accounts of entanglement have been defended to and fro in the literature (Teller, 1986; Esfeld, 2004; Morganti, 2009), and are now considered to be part and parcel of the general framework of ontic structural realism (Ladyman, 1998; French, 2014; McKenzie, 2017). In particular, Cinti et al. (2022) aims to add some relevant details to such a view, namely, to make explicit what the relata of the entanglement relation are, how many entanglement relations there are once the relata are fixed and what parts of the mathematical formulation of the theory represent these metaphysical relations. In Cinti et al. (2022), we argue that the relata of the entanglement relation are quantum degrees of freedom (DoFs),[7] i.e. physical properties, and that such a relation is unique given a set of DoFs. Moreover, we argue that such an entanglement relation must not be read out of quantum states.[8] Rather, we contend, contra standard definitions of entanglement, that *mutual information* represents the entanglement relation.[9] Note that, for the present thesis, the *mutual information* and *entanglement entropy* can be equivalently taken to be representative of entanglement relations since we are dealing at most with a composite system of two quantum systems (such as the black hole and its radiation). In this case, the *mutual information* and *entanglement entropy* coincide. In Cinti et al. (2022), we decided to consider *mutual information* instead of *entanglement entropy* because we wanted to have a unified framework to describe bipartite and multipartite systems.

[7]To be more precise, entanglement is a relation between dynamical DoFs, i.e. those that undergo time evolution. This fact excludes from the province of entanglement non-dynamical quantities such as, for example, superselected ones.

[8]As such, our account opens the perspective of entanglement realism also to those who are antirealists on the quantum state, i.e. those who consider the quantum state as a mathematical tool to calculate probability measures for measurement outcomes but still want a realist ontology of QM (Rovelli, 2016, 2018).

[9]Let me define what *mutual information* is in the context of quantum theory. For a precise treatment of mutual information and how it can be thought to define entanglement relations, see Cinti et al. (2022). Given two system, a natural way to represent entanglement basis-independently starts with the *von Neumann entropy*:

$$S(\rho) = -tr(\rho \ln \rho) \ , \tag{4.1}$$

where ρ is a density matrix. One can immediately see that $S(\rho)$ is basis-independent since $S(\rho) = S(U^\dagger \rho U)$, where U is an arbitrary change of basis. Then, to characterise entanglement for bipartite systems, we use the mutual information:

$$I(A, B) = S(\rho_A) + S(\rho_B) - S(\rho_{AB}) \ , \tag{4.2}$$

where ρ_A and ρ_B are the reduced density matrices of the subsystems A and B. This definition can be straightforwardly extended to multipartite systems.

Let us now summarise some further properties of the entanglement relation as we characterise it in Cinti et al. (2022):

Externality/Basis-Independence: the entanglement relation as characterised by the mutual information is an external relation

Relata: we take the relata of the entanglement relation to be the degrees of freedom (DoFs) of quantum systems. Indeed, note that the Hilbert space of any composite system is, in general, given by the tensor product of the DoFs involved. Thus, it seems natural to take the DoFs as the relata.

Adicity: the entanglement relation is an $N-$place relation, where N is the number of DoFs, i.e. it is a multigrade relation (Leonard and Goodman, 1940; Oliver and Smiley, 2004).

Non-transitivity: for the entanglement relation, as represented by the mutual information, transitivity does not make sense insofar as the entanglement relation is not internally ordered.[10]

Non-supervenience: since the entanglement relation, as represented by the mutual information, is an external relation, it does not supervene on the intrinsic properties of its relata, i.e. the DoFs. Thus, the entanglement relation is non-supervenient.

To conclude, in Cinti et al. (2022), we argue that entanglement is a unique metaphysical relation characterised by the mutual information.[11] Entanglement is a non-supervenient external multigrade relation whose relata are N DoFs. After fixing the number of entangled systems, the relation is unique and not relative to any specific observable. Such a relation is characterised basis-independently by the mutual information (or the entanglement entropy for bipartite systems) and is non-transitive and irreflexive. Finally, it turns out that it is also positive, symmetric and monotonic.

[10]Let me expand on what "internally ordered" means in this framework. The entanglement relation, as understood in Cinti et al. (2022), is a $N-$place relation involving all the DoFs of a system. Hence, one cannot claim that entanglement is transitive because when three quantum systems A, B and C are entangled, there are no internal relations such that 'A is entangled with B', 'B is entangled with C' and 'C is entangled with A'; rather, A, B and C partake into a unique relation of entanglement.

[11]Recall the discussion above for the present purposes, mutual information and entanglement entropy can be used interchangeably.

4.3.2 The Entanglement Causal Structure

As we have seen in the previous section, entanglement can be interpreted as an external relation between distinct quantum systems. We can thus, applying the definition of Causal Structure given above, define the *Entanglement Causal Structure*, in which the theory T is Quantum Mechanics (or, better, Quantum Field Theory), the set of objects Ω are quantum systems, and the relation R is the entanglement relation R_{ME} of *being (maximally) entangled with*, with all the features described in Cinti et al. (2022).

The correlations involved in entangled states are particularly robust because they are not merely accidental correlations between two subsystems but are codified in the laws of nature of QM. They thus have, insofar as we take those laws to be a reliable guide at least to nomological modality, an important modal, and thus counterfactual, robustness.[12] Indeed, such counterfactual robustness of entanglement correlations has convinced some philosophers that we should understand these correlations as causal dependencies (Maudlin, 2002). For simplicity of exposition, we refer to entanglement correlations with talk of causal-like dependencies. Note, however, that this is not necessary for the definition of Causal Structures, as also discussed in §4.1. The robustness of the correlations is sufficient to use the entanglement relation as a possible relation R in the definition of Causal Structures. In particular, nothing in our arguments hinges on some of the features typical of stronger accounts of causation, be they either energy-momentum exchange among causally related systems or causally significant interventions being possible. Since talking in causal terms renders the discussion much more straightforward, we make use of this talk. Those who disagree that entanglement correlations can be understood as encoding causal dependencies are free to substitute causality with robust counterfactual correlations. Indeed, outside of issues of linguistic simplicity, we encourage the reader to think, for the remainder of the thesis, of all talk of causality as being concerned only with counterfactually robust correlations and not necessarily with any stronger notion of causation.

Moreover, note that the causal structure of entanglement does not need to capture, in any interesting way, the fundamental causal structure. This would imply the stronger condition of entanglement fundamentalism (Jaksland, 2020). We only need it to capture the structure of entanglement relations among quantum systems encoded in the

[12] Here by robustness we only refer to the counterfactual stability of entanglement correlations. We are not referring to the sense of robustness developed by Redhead (1987) as a way to examine the stability of entanglement.

robust counterfactual connection between them.

Finally, one might worry that, given our definitions, vacuum entanglement in QFT could already lead to significant departures from the causal structure of GR without any need for QG. However, the notion of causal structure that we are employing only involves counterfactual connections. As such, it does not require the existence of signals propagating between two systems so connected. Since entanglement only involves counterfactual connections (Maudlin, 2011) but not signal propagation (by the no-signalling theorem), the causal structure of entanglement in QFT does not necessitate revision to the causal structure of spacetime. In QFT, one can superimpose the causal structure of entanglement with the causal structure of spacetime. This is because the causal structure of spacetime is only sensitive to signals' propagation, and entanglement's causal structure does not rely on signal propagation. QG is required since, in QG, it is spacetime that manifests entanglement, and thus we need a way to reconcile these two conflicting causal structures.

4.3.3 Are Relativistic and Entanglement Causal Structure Compatible?

Note that this way of recasting connections between entities in different theories can be handy in studying the possible interplay between them. In particular, we can analyse the discussions about the supposed non-local character of entanglement Jarrett (1984); Maudlin (2002). This feature is traditionally associated with having spacelike separated spacetime points (or regions, depending on our understanding of location for quantum systems) at which objects are in entangled states. There are points causally separated for the causal structure of spacetime (defined by R_{LC}) but causally related for the causal structure defined by the entanglement relations R_E. In other words, in the language of causal structures, this means that for two quantum systems A and B, it can be the case that $R_{ME}(A, B)$ while $\neg R_{LC}(A, B)$.[13] This only holds for entanglement between spacelike separated systems. From the present perspective, we ignore the case of timelike

[13] Observe that our discussion has been cast in the context of relativistic spacetime, i.e. the relation of causal connectedness in **(CS)** is given by R_{LC}. While this assumption is useful in the context of my thesis since black holes are intrinsically relativistic objects, it is by no means required by any theory considered. Indeed one could recast the problem of (in)compatibility between the two causal structures in the context of Galilean spacetime, by using as the relation of causal connectedness the relation R_G of *being in spatial contact with*, and the arguments would run parallel to those of this section.

entanglement (Olson and Ralph, 2012), which is not relevant to the present discussion since the two systems can be connected by a causal curve, explaining their correlation. Note that this fact does not imply action-at-a-distance since entanglement, and thus the relation R_E, only imply the existence of robust counterfactual connections which, however, do not necessitate any form of action-at-a-distance, as argued, for example, in Myrvold (2015). Indeed action-at-a-distance, as (Wallace, 2012, p. 292) notes, would be incompatible with relativity.[14]

Their being causally separated for the causal structure of spacetime but causally related for the entanglement causal structure has been the source of many conceptual puzzles. However, results such as the no signalling theorem and discussions such as that of Jarrett (1984) have convinced philosophers and physicists alike that entanglement cannot violate locality in any observable way. Since entanglement does not violate locality, one might expect that the two causal structures can be combined by superimposing them without modifying one or the other.

As we will see in the following chapters, what AMPS and AMPSS show is that this expectation cannot survive in the context of quantum black holes. While the single entanglement connections between pairs of spacelike separated quantum systems are not problematic, the overall pattern of spacetime and entanglement connections remains incompatible. Indeed, one can generate situations, such as that of AMPS, where there are three or more distinct[15] maximally entangled quantum systems, something which should not be possible according to the rules of quantum entanglement. Entanglement and spacetime causal structures are incompatible, despite the careful considerations that led some to think they were. We lacked the theoretical tools to capture the consequences of this incompatibility, theoretical tools that were finally developed in the study of quantum black holes and put to use by AMPS. Understanding the extent of this incompatibility, and showing how to overcome it, is the task of resolving the firewall paradox.

[14] Note that, one could use a different terminology, namely that counterfactual dependence is enough to have a form of action at a distance, not mediated by a signal. One, for instance, could take action at a distance to be simply what's established by the violation of Bell-type inequalities. In this sense, an action at a distance that doesn't require any signal propagating between the spacelike separated subsystems. In our discussion, however, we are taking, together with many philosophers, such as (Wallace, 2012, p. 292), action at a distance to require a propagating signals.

[15] Which, as we have observed at the beginning of this section, is justified in terms of their being spacelike separated.

Chapter 5

Black Hole Paradoxes: a Philosopher's Take

This chapter aims to provide a philosophical discussion of two black hole paradoxes: the black hole monogamy paradox, or AMPS paradox, discussed in section §3.3, and the interior operator paradox, or AMPSS paradox, discussed in section §3.4. The discussion will be done in two different steps:

First of all, we are going to analyse the underlying principles that are responsible for each paradox. In doing so, we will point out the presence, in both paradoxes, of an implicit assumption, which is hidden in the usual description of the paradoxes given by physicists. My first contribution is pointing out precisely the principles that enter into conflict during black hole evaporation.

Second of all, we are going to argue which are the possible strategies to avoid the paradoxes, i.e. which are the principles we could drop to obtain a non-paradoxical description of black hole evaporation. In doing so, we will argue which are the logically admissible solutions and match them with strategies that physicists have already used in the literature on black hole paradoxes. We are also pointing out which solutions can be considered more conservative, i.e. which solutions allow us not to renounce to fundamental physical principles of our best theories.

This work interests physicists and philosophers researching on the foundations of black holes and QG. The conceptual insights provided by our analysis both serve to elucidate the philosophical foundations of the topic and help researchers in this field better appreciate the tools they are using. This way, we bridge the gap between the two communities, furthering the study of black holes as both a technical and conceptual topic. As argued in §2.4, I believe this is one of the most important ways in which the philosophy of physics can contribute to cutting-edge research in physics.

In providing a philosophical description of the two black hole paradoxes, we will see that modern solutions proposed for both paradoxes rely on the same conceptual strategy. This can be considered a result of interest for physicists. Indeed, arguing that the resolution of the AMPS and the AMPSS paradox are two instances of the same conceptual move was not evident from the physics literature itself. From our philosophical reconstruction, we can see how the two resolutions are indeed connected and are instances of the same conceptual move, which, as we will see, is connected with the ER=EPR conjecture. Moreover, the way we will explore the paradoxes will be significant for studying the ontology of black holes in quantum gravity. Indeed, we can analyse the ontology of black holes from the point of view of the various resolutions of the paradoxes, which rely on dropping a particular physical principle. The implication of the multiple resolutions of black hole paradoxes on the ontology of black holes will be the content of chapter §6.

The section is structured as follows: in §5.1, we will analyse the AMPS paradox from the point of view of the causal structures; in §5.2, we will explore the AMPSS paradox and its possible resolutions.

5.1 AMPS

As we already discussed throughout this thesis, philosophers did not ignore the importance of the conceptual study of black holes. Works of this type are Wallace (2020b) and Belot et al. (1999a). However, both articles deal with variations of the black hole information loss paradox, discussed in section §3.1, and the Page Time paradox, discussed in section §3.2, which revolve around Hawking's original idea that black hole physics might be non-unitary (and ways to avoid this conclusion). Nevertheless, most contemporary high-energy physicists are not usually concerned with the unitarity of black hole physics (which, especially among string theorists, is taken to follow from the AdS/CFT duality Maldacena (1999c); Ammon and Erdmenger (2015), where unitarity is a standard feature of the boundary CFT), but rather with the structure of the interior of the black hole. In this regard, the *AMPS paradox* plays a central role. Our goal as philosophers of physics is to study the conceptual foundations of the firewall paradox and to explore how dropping an implicit assumption on the structure of spacetime, what we call *spacetime distinctness*, resolves it. In particular, we highlight, by looking at concrete physical examples, how recent discussions regarding the resolution of the firewall paradox Papadodimas and Raju (2013); Maldacena and Susskind (2013); Papadodimas and Raju (2016); Hayden and Penington (2019a); Almheiri et al. (2019c); Penington (2019); Almheiri et al. (2020c,c) appear to rely crucially on this strategy.

To clarify the AMPS paradox's conceptual structure and resolution, we rely on the notion of *causal structures*, which we developed in §4.1. The upshot of our analysis is that the firewall paradox crucially depends on the assumption that relativistic locality, i.e. that only causal curves can carry causal influences, broadly understood in terms of counterfactually robust correlations, is preserved in QG and that a natural way to resolve the paradox is to drop this assumption.

The section is structured as follows: in §5.1.1, we give a conceptually oriented presentation of the firewall paradox and study it in terms of causal structures. In §5.1.2 and §5.1.3, we study some concrete physical models connected to the ER=EPR conjecture and at the core of recent discussions of the firewall paradox and clarify that they solve the paradox precisely by dropping the implicit assumptions which we called **spacetime distinctness**. §5.1.4 then concludes.

5.1.1 AMPS and Causal Structures

Recall from section §3.3 that the AMPS paradox is a paradoxical situation occurring during the evaporation process of a black hole at the Page time, approximately half of the lifetime of a Schwarzschild black hole. AMPS pointed out that, to have a consistent description of black holes in semiclassical gravity, compatible with both Quantum Mechanics and General Relativity, one arrives in a situation in which a late radiation mode L should be entangled *at the same time* with both an early radiation mode E and an interior mode B, violating the monogamy of entanglement, which is a fundamental principle of Quantum Mechanics. Let us now clarify where the tension lies in the AMPS argument.

Beyond the four postulates listed in §3.3.1, the fact that B and E are separate systems plays a fundamental role in the paradox. To see why this is crucial for deriving the paradox, let us imagine that B and E were not distinct systems. In this case, we do not have any paradox since L would be entangled with just one system, i.e. the elementary system composed of B and E. For black holes, the distinctness between B and E is justified because B lies behind the horizon of the black hole. B is thus, within a fixed spatial hypersurface at time $t > t_p$, causally isolated from L and E. Recall that we must study distinctness between B and E on a spatial hypersurface at time $t > t_p$. Indeed, we need to understand if monogamy is violated after t_p, and monogamy concerns entanglement among quantum systems *at a given time*. Thus, we are interested in the entanglement relations between quantum systems on a spatial hypersurface at a specific time $t > t_p$. And, on such spatial hypersurface, the exterior and interior of the black hole are spacelike separated.[1] However, this is not an assumption as innocent as it might sound: it is equivalent to the claim that relativistic notions of separability and locality, based on spacelike and causal connections, are retained in the regime of QG. Let us focus on this point, as it will be crucial in the rest of this section.

As we have mentioned, to ensure the violation of monogamy, and thus the firewall paradox, one has to regard B and E as distinct systems, where by distinct we mean that their degrees of freedom are independent. This fact follows from monogamy of entanglement, giving an upper bound on the number of independent degrees of freedom with which a certain quantum system, in our case L, can be entangled. Indeed, (i) we know that B and E are spacelike related since, at fixed t, one is in the interior and the other in the exterior of the black hole. Furthermore, (ii) from the study of QFT (and in particular of Algebraic QFT[2]) we have come to accept that if two algebras of observables[3] are mutually commutative, then they represent two distinct systems,[4] since their degrees of freedom are completely independent when this is the case. From the axiom of *microcausality*, (iii) we have that the algebras of observables connected to two spacelike related regions must commute, i.e. $[\mathfrak{N}(\mathcal{X}), \mathfrak{N}(\mathcal{X}')] = 0$, where $\mathfrak{N}(\mathcal{X})$ and $\mathfrak{N}(\mathcal{X}')$ are the algebras of observables associated to two spacelike related regions $\mathcal{X}$ and $\mathcal{X}'$. Thus, from (i)-(ii)-(iii), the two systems must be distinct.

Note, however, that the microcausality axiom encodes the locality properties of classical relativistic spacetime since it relies explicitly on the notion of spacelike separation,

[1] Even if one were to consider an observer who takes E and L and jumps into the black hole meeting B, thus making E and B, not spacelike separated, it is still the case that the distinctness of E and B, and thus the AMPS paradox, is grounded in their being spacelike separated before the observer jumping in. Thus, in what follows, we will not consider this situation and stick to the more general formulation of the paradox outlined above.

[2] For reviews of AQFT see Halvorson (2007); Haag (1996).

[3] Note that its algebra of observables identifies a system.

[4] Though see Earman and Valente (2014) for some subtleties about this claim.

which we have no guarantee will be retained at the level of quantum spacetime. Thus we have to *assume* that this notion of locality, developed for relativistic spacetimes, can be extended seamlessly beyond GR. If this were not the case, then B and E might not be distinct at the quantum level, opening the door to resolving the paradox by observing that there is no violation of monogamy since B and E are not distinct systems which implies that their degrees of freedom are not independent and that L is entangled with fewer degrees of freedom than those manifest in the semiclassical description (more on this in §5).

We have thus seen that to the four postulates above, AMPS (implicitly) add a fifth one:

(SD) **Spacetime Distinctness:** spacelike separated systems are distinct, i.e. mutually commuting.

Note, again, that **(SD)** encompasses the axiom of microcausality of algebraic QFT, which is based on the relativistic notion of locality since it relies explicitly on spacelike separation. However, we have no guarantee that such a notion of locality holds in QG since spacetime might have different properties when considering quantum effects. In particular, it might happen that the algebras of two spacelike-separated quantum systems do not commute, violating **(SD)**. Indeed, this kind of phenomenon is a possible way to elude the AMPS paradox. Let us then discuss firewalls.

With this in mind, we can then reformulate the paradoxical conclusion of AMPS as follows: in a black hole spacetime, we have three subsystems L, B and E, such that B and L are distinct systems, i.e. $\mathfrak{N}(B)$ commutes with $\mathfrak{N}(L)$, and they violate the monogamy principle, i.e. L is maximally entangled with both B and E.

To better keep track of the various moving parts of the AMPS argument, and to elucidate its conceptual content, let us recall the definition of causal structure:

(CS) **Causal Structure**: given a theory T, we say that the causal structure according to the theory T is given by a set of spacetime regions/objects (with their physical state) and a relation R which determines if two objects/regions of spacetime can or cannot be causally related.

To recast the AMPS paradox in terms of causal structures, let us take quantum systems as our objects and consider two different relations, which define two different causal structures. One is the relation R_{LC} of *being connectable by a causal curve* that we have encountered in §4.2, which defines what we call the causal structure of spacetime. The other is the relation R_{ME} of *being maximally entangled*, which defines what we call the causal structure of entanglement, defined in §4.3. The core claim of AMPS is then that, under the four postulates detailed in §3.3.1, there are three quantum systems L, B and E such that $R_{ME}(L, B)$ and $R_{ME}(L, E)$, and B and E are distinct, which follows from $\neg R_{LC}(B, E)$, i.e. B and E are spacelike related. Our four assumptions (supplemented with **spacetime distinctness**) are then in violation of the monogamy of entanglement (see Figure 3.5).

This is a simple example in which we understand the incompatibility between the causal structure of spacetime and entanglement, which we mentioned at the end of chapter §4.3. Indeed, we have seen that combining the causal structure of spacetime and entanglement by simply superimposing them leads to a contradiction. Thus, one must consider possible interactions between the two causal structures to solve the problem. Indeed, this could have been argued long before the black hole evaporation process because we do not expect spacetime to have the same features as those described by the causal structure of spacetime at the quantum level.

To resolve this paradox, one has three logically admissible possibilities:

(i) accept $\neg R_{ME}(L, B)$, i.e. L and B are not entangled. This is AMPS' answer and implies a firewall on the horizon.

(ii) accept $\neg R_{ME}(L, E)$, i.e. L and E are not entangled. This answer is equivalent to Hawking's calculation and implies the non-unitarity of black hole evaporation.

(iii) accept that B and E are not distinct. This answer implies as we see in §5, the modification of the causal structure of spacetime.

Note that there is a fourth possibility that we do not consider in this section, which is dropping $\neg R_{LC}(B, E)$. We do not consider this possibility because this would imply a massive violation of General Relativity. However, note that solutions which could be understood as instances of this very extreme move had been presented in the physics literature (Giddings, 2013b,a). Indeed, Giddings's approach implies the existence of non-local interactions at the level of semiclassical, effective field theory. Thus, Giddings's proposal would imply the acceptance of $R_{LC}(B, E)$, i.e. B and E being causally connectable from the point of view of the causal structure of spacetime presented in §4.2. As a matter of fact, as observed in Harlow (2016), Giddings's proposal is better characterised as a type of firewall. However, from our classification, it appears to be somewhat different from the firewall proposed in AMPS. Moreover, Giddings's proposal seems to imply either the violation of the laws of black hole thermodynamics (Almheiri et al., 2013) or modifications of the Schwarzschild geometry away from the horizon Giddings (2013c, 2014), both consequences which are not shared by (iii).

The solution that we study in this section is (iii).[5] Please observe that the causal structure formulation allows us to clarify the dependence of the paradox on a precise notion of when two systems are distinct, something that is not evident from the formulation of AMPS. Furthermore, this way of expressing the paradox allows us to connect explicitly with previous discussions on the supposed non-local character of entanglement Jarrett (1984); Maudlin (2002). This feature is traditionally associated with having spacelike separated systems which are, however, entangled. In the language of causal structures, this means that for two quantum systems A and B, it can be the case that $R_{ME}(A,B)$ while $\neg R_{LC}(A,B)$. What AMPS show is that the peaceful coexistence argued by Jarrett (1984) cannot survive in the context of quantum black holes. While the single entanglement connections between pairs of spacelike separated quantum systems are not problematic, the overall pattern of spacetime and entanglement connections remains incompatible. Indeed, one can generate situations, such as that of AMPS, where there are three or more distinct[6] maximally entangled quantum systems, something which should not be possible according to the rules of quantum entanglement.

Entanglement and spacetime structures are incompatible, despite the careful considerations that led some to think they were. We lacked the theoretical tools to capture the consequences of this incompatibility, theoretical tools that were finally developed in the study of quantum black holes and put to use by AMPS. Understanding the extent of this incompatibility, and showing how to overcome it, is the task of any resolution of the firewall paradox.

5.1.2 No Firewall on the Horizon

In this section, we analyse the resolution of the firewall paradox associated with the third strategy, which we dub route (iii). Before delving deeper into this topic, however, it is useful to recall our basic setup. First of all, we have begun in §3.1 and §3.2 by

[5] In section §5.2.1, I provide a defence on why I think that route (iii) is the more promising solution to the AMPS paradox.

[6] Which, as we have observed at the beginning of this section, is justified in terms of their being spacelike separated.

reviewing the two central paradoxes in black hole physics: the black hole information paradox and the Page time paradox. In particular, the Page time paradox shows the incompatibility of unitary evaporation, as encoded in the Page curve of the black hole entropy, with Hawking's original description of black hole radiation Hawking (1976). We have also seen that AdS/CFT has been argued to show that unitary evaporation is ultimately the correct option. However, as shown in Almheiri et al. (2013), unitarity is still not sufficient to give a consistent description of black holes and, in particular, of the interior. Indeed, one runs into the *firewall paradox* (§3.3): from four reasonable postulates, one can derive a situation violating the monogamy of entanglement. Furthermore, we have proposed to study the conceptual structure of the firewall paradox in terms of *causal structures* (§4.1). In this framework, the paradox arises as an incompatibility between entanglement and spacetime causal structures. In particular, we have seen that a fundamental ingredient for the paradox to arise is **(SD)**, which is equivalent to the claim that the spacetime causal structure of GR is imported in the final theory of QG.

While AMPS have suggested that one should resolve the paradox by positing that the infalling observer encounters a firewall of high energy quanta at the horizon, retaining **(SD)** at the cost of the equivalence principle, such a way to resolve the paradox is not the only one possible. In particular, we have argued in §5.1.1 that dropping **(SD)** is a viable approach to resolving the paradox. In this section, we are going to introduce some concrete physical models which are at the centre of recent discussions among physicists regarding the resolution of the firewall paradox Papadodimas and Raju (2013); Maldacena and Susskind (2013); Papadodimas and Raju (2016); Hayden and Penington (2019a); Almheiri et al. (2019c); Penington (2019); Almheiri et al. (2020c,c) and argue that they avoid the paradox by instantiating the violation of **(SD)**. This analysis serves to show that these various approaches all share the same basic conceptual strategy and to better appreciate its implications. These models have been constructed in the context of AdS/CFT[7] and, in particular, they are instances of the ER=EPR conjecture Maldacena and Susskind (2013).[8] However, let us stress that these models do not strictly depend on the conjecture's truth but rather provide evidence for it since

[7]Note that, in what follows, we only use AdS/CFT as a mathematical tool to construct specific physical models. In particular, we do not discuss any of the philosophical issues connected to holographic dualities De Haro et al. (2016).

[8]Here, ER stands for Einstein and Rosen, from the seminal article Einstein and Rosen (1935) introducing wormholes. In contrast, EPR stands for Einstein, Podolsky, and Rosen from the article Einstein et al. (1935), which first pointed out the EPR paradox and the non-local character of entanglement.

they can be constructed (somewhat) independently of ER=EPR. Ultimately we will see that the non-local connections characteristic of the ER=EPR conjecture, and concretely instantiated in these models, engineer the violation of **(SD)** in which we are interested. Thus how ER=EPR avoids, the firewall paradox is a paradigmatic instance of route (iii)'s strategy for resolving the paradox. As the reader is sanguine regarding the prospects of ER=EPR, our discussion provides evidence that the paradox's solution lies in the violation of **(SD)**. For the reader to be more sceptical about the prospects of these approaches, we show that the violation of **(SD)** resolves the paradox in certain specific models and clarifies what this resolution implies while leaving open if the same strategy applies to more general cases.

We proceed in §5.1.3 by discussing the instructive, though not realistic, case of the eternal AdS black hole where we can see how ER=EPR undermines **(SD)** fundamental to the AMPS paradox. In §5.1.4, we then move to the case directly relevant to the AMPS paradox, that of an evaporating black hole formed from gravitational collapse.[9]

5.1.3 Eternal Black Holes

Let us start from the case of eternal[10] AdS black holes, following the treatment of Maldacena and Susskind (2013). While these are not the evaporating black holes generated from gravitational collapse, to which the AMPS paradox applies, they can still be instrumental in testing and developing ideas regarding black holes' structure. In particular, the purpose of starting from the case of a two-sided black hole[11] is that it serves a useful pedagogical role since it allows to phrase various questions regarding the interior structure of black holes to a level of precision hard to attain in the context of standard, evaporating, one-sided black holes.[12] It is thus useful to start from this most

[9]Further constructions, which implement the same idea and apply to generic black holes in AdS, are presented in Papadodimas and Raju (2016); Almheiri (2018). Note that these constructions crucially rely on the notion of *state-dependence*, originally introduced in Papadodimas and Raju (2013).

[10]By eternal here, we mean a black hole which has always existed and will always exist. Thus, the black hole has not formed via gravitational collapse and is not subject to evaporation. In AdS, the non-evaporation of the black hole is due to the reflecting boundary conditions of the spacetime, which mean that the Hawking radiation emitted from the black hole bounces back inside the black hole upon reaching the boundary of AdS spacetime. Thus, the black hole is in equilibrium with exterior spacetime and does not evaporate. To make an AdS black hole evaporate, one needs to have absorbing boundary conditions, which we discuss in §5.1.4.

[11]By a two-sided black hole, we mean a physical system with two event horizons. As such, a two-sided black hole has two exterior regions, usually called the *left* and the *right* exterior. Eternal AdS black holes are two-sided black holes.

[12]By a one-sided black hole we mean a physical system with one event horizon. As such, a one-sided black hole has only one exterior region.

basic case.

Eternal AdS black holes are holographically dual to a couple of entangled CFTs (called the *left* and *right* CFT depending on which of the two exterior regions of the AdS black hole they describe) in the so-called *thermofield double* state:

$$|\psi\rangle = \sum_j e^{-\beta E_j/2} |j\rangle_L \otimes |j\rangle_R \tag{5.1}$$

where E_j are the energy levels of the CFT and $\beta = 1/T$, where T is the temperature of the black hole, $|j\rangle_L$ and $|j\rangle_R$ are states in the left and right CFT. The core of ER=EPR is the conjectured equivalence between entangled systems and wormhole geometries, i.e. that between any two entangled quantum systems, there is a wormhole, possibly of Planckian size.[13] The eternal AdS black hole has a dual interpretation: we can either understand it as a system made of two entangled black holes or as two black holes connected by a wormhole. This dual way of looking at eternal AdS black holes is at the heart of the ER=EPR proposal. Indeed, the eternal AdS black hole is a particularly special case of the ER=EPR conjecture in the sense that we do not need to modify the classical geometry of the black hole to get ER=EPR. As it were, the wormhole is already there in the eternal AdS black hole. In particular, no quantum wormhole is needed to verify the conjecture in this case, only classical geometry, making the subsequent discussion much more accessible. The primary lessons that we learn in this case, however, carry over also to the more realistic cases that we treat later, where no such convenient semiclassical picture is available.

The semiclassical features of the eternal AdS black hole are at the heart of how ER=EPR solves the AMPS paradox. To see why let us start by constructing a firewall-like situation in the context of the eternal AdS black hole (Figure 5.1). We start at time t by taking a pair of entangled qubits A and B, on the right side of the two-sided eternal black hole, with A behind the black hole horizon and B outside the horizon. Let us now apply a CPT transformation[14] to A. Since this transformation is a symmetry, sends t into $-t$, and exchanges left and right, we get a qubit A', equivalent to A, at time $-t$ in the left exterior region of the black hole. Furthermore, it is clear from the

[13]Susskind distinguishes between a modest and an ambitious version of the conjecture Susskind (2016b). Modest ER=EPR is supposed to apply only to entangled black holes, while ambitious ER=EPR applies to any entangled systems.

[14]CPT here stands for *charge*, *parity* and *time reversal*. CPT transformations are among the fundamental symmetries of quantum systems, such as the one we are considering here.

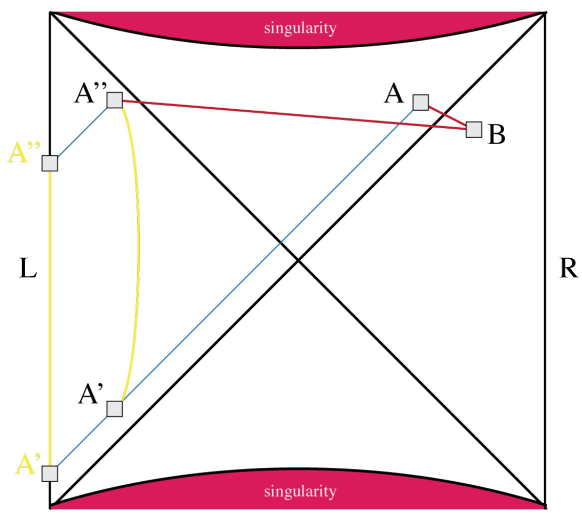


Figure 5.1: The Penrose diagram of an eternal AdS black hole (with L the left exterior and R the right exterior) with an AMPS situation and its ER=EPR resolution. Here red lines represent entanglement and blue lines semiclassical bulk evolution, while yellow lines represent boundary evolution and its bulk dual. Furthermore, A is an interior mode, A' is its CPT transform (with its holographic dual in yellow), and A'' is a mode on the left horizon (with its holographic dual in yellow) obtained by evolving A'. The AMPS-like situation comes from the maximal entanglement between A'' and B and A and B. The resolution is given by the existence of a unitary connection between A'' and A.

Penrose diagram that we can evolve A' into A with the bulk equations of motion. Let us write this as $A' \to A$. In particular, A' is entangled with B since we just applied a CPT transformation to A, and A is entangled with B. Since A is a qubit, it has three components A_i where $i = 1, 2, 3$, which are Pauli matrices. We thus have that:

$$[A_i, A_j] = i\epsilon_{ijk}A_k \neq 0 \tag{5.2}$$

Since, however, by forwards time evolution with the bulk equations of motion, we have that A' evolves in A, we can also write:

$$[A'_i, A_j] \neq 0 \tag{5.3}$$

To arrive at this result, we have relied on the fact that we can view the eternal AdS black hole as a wormhole connecting two horizons. A crucial step was the forwards time evolution of A' in A, relying on A' passing through the wormhole, as evident in Figure 5.1. However, we also know that we can regard the eternal AdS black hole as two entangled black holes, which means that we can evolve qubits in the left exterior (dual to the left CFT) independently from the right exterior (dual to the right CFT) since these are just two disconnected spacetimes. In particular, if we evolve A' forwards in time up to time t, with the left CFT Hamiltonian, we obtain the qubit A'', which is naturally understood as the holographic dual to a qubit living on the left horizon.[15] Since A'' is just the product of evolving forwards in time A', it carries the same information (in the sense that they are related by a unitary operator, the left CFT Hamiltonian). In particular, it is entangled with B, giving us a firewall-like situation. At t, we have a qubit B entangled with a qubit in the interior (A) and a distant, far away qubit (A''). Here B is equivalent to L in our formulation of AMPS while A is equivalent to B, and A'' to E (remember Figure 3.5). It would thus seem that we have, here too, a violation of the monogamy of entanglement. However, we can immediately see the resolution of this apparent paradox. Since A'' follows from applying the left CFT Hamiltonian to A', and $A' \to A$, we can also write $A'' \to A$.[16] But then we can substitute A'' for A' in (5.3), giving us:

$$[A_i'', A_j] \neq 0 \tag{5.4}$$

(5.4) resolves the apparent violation of monogamy that we have engineered since it tells us that A and A'' are not independent qubits (since they do not commute). As such, it is not the case that B is maximally entangled with two different systems, leading to a violation of the monogamy of entanglement. Since A and A'' are not distinct systems, B is entangled with only one system, represented by both A and A'', in a fundamentally non-local manner. The mistake in the reasoning which led us to an apparent violation of monogamy was the (background) assumption of **(SD)**, which told us that two spacelike separated systems must commute and thus be distinct. Since A and A'' are spacelike separated, it was only natural to assume that they were two different, distinct systems. What our analysis shows is that, in the context of the eternal AdS black hole where

[15] To be precise, the left stretched horizon, though this is not relevant to the present discussion.

[16] Intuitively, we can understand this as first evolving backwards in time from A'' to A' with the left CFT Hamiltonian, and then forwards in time from A' to A with the bulk equations of motion.

ER=EPR is already a feature of the semiclassical geometry, the assumption of **(SD)** falls apart, which is the essence of (iii)'s resolution of the AMPS paradox. Even spacelike separated, distant objects can still nonetheless depend on each other, and thus not be distinct objects after all.[17]

This analysis is best understood in the language of causal structures **(CS)**. In §5.1.1, we have seen that the monogamy paradox can be recast as the mismatch between the causal structure of spacetime and entanglement. Since $R_{ME}(A, B)$ and $R_{ME}(A'', B)$ while $\neg R_{LC}(A'', A)$, which, by **(SD)**, implies that A'' and A are distinct systems, we come to the conclusion that B is maximally entangled with two distinct systems, violating monogamy. The way ER=EPR allows us to resolve the paradox, in the context of the eternal AdS black hole, is by defining a more general causal structure, characterised by the relation R_{WH} of *being non trivially connected*, which obtains if and only if two entities A and B can non trivially influence each other. A non-trivial influence is manifested by the presence of counterfactually robust correlations between A and B. We call this the *generalised causal structure*.

Observe that it is always the case that we can embed the causal structure of spacetime and entanglement in the generalised causal structure by observing that both R_{ME} and R_{LC} are supposed to produce robust counterfactual correlations. However, and here lies causal structures' usefulness in understanding the ER=EPR resolution of AMPS, in certain situations, such an embedding is not an isomorphism, i.e. there are systems connected by R_{WH} which are connected neither by R_{ME} nor R_{LC}. Indeed, in the eternal black hole in AdS, the wormhole connects the left and right exterior, making it possible to have a non-trivial connection between A and A'', despite their not being entangled and being spacelike related. Indeed, it is this non-trivial connection, captured by R_{WH}, which leads to the violation of **(SD)**, as per (iii).

The eternal AdS black hole serves as a simple motivating case to understand how violations of **(SD)** naturally emerge in studying black holes. In the next section, we can now move on to the study of evaporating black holes formed from gravitational collapse. Let us, however, remark once more the most critical intuition underlying the ER=EPR conjecture: it is the idea that the overall structure of spacetime, as it emerges from QG, is much more complicated than its *naive* semiclassical description would lead us to

[17] It would be an interesting project to understand better how the notions of distinctness and dependence that we use here can be related to the notions of distinctness and dependence analysed in Schaffer (2016) in the framework of metaphysical grounding. Such a project, however, goes beyond the scope of the present thesis.

believe. There are many more connections that are not accounted for by merely thinking in terms of causal curves, and these connections are central to a proper understanding of black holes. Furthermore, while in some lucky cases (such as the eternal AdS black hole), we can understand these connections in geometrical terms as wormhole geometries, this is not possible in other cases. Instead, we have to resort to the more general idea of there being connections absent in the semiclassical description, imaginatively called *planckian* or *quantum* wormholes in Maldacena and Susskind (2013). These non-local (from the perspective of semiclassical spacetime) connections are the heart of the ER=EPR conjecture and lie at the core of the constructions described here.

5.1.4 Evaporating Black Holes from Gravitational Collapse

In this section, we study the firewall paradox in a situation where the black hole is formed from gravitational collapse and can evaporate, a problem studied extensively in Penington (2019). Even if we already have seen some of the features of black hole evaporation in AdS/CFT-inspired setup in §3.4.2, let me give an in-depth analysis of the physics behind it in this section to appreciate the fact that the solution to the AMPS paradox proposed in Penington (2019) is an instance of route (iii), in the very same way the Maldacena and Susskind (2013) was for the eternal black hole case.

As we already said, the black hole in AdS spacetime considered in §5.1.3 is not evaporating since it has reflecting boundary conditions and cannot give a real example of the firewall paradox. Indeed, the conformal boundary of AdS reflects the Hawking quanta into the black hole, reaching thermal equilibrium since the number of emitted and reflected quanta compensate (see Figure 5.2).

One strategy to build an evaporating black hole is to take a black hole formed from collapse and place it into a spacetime whose boundary is not completely reflecting, i.e. in a spacetime that permits some Hawking quanta to escape outside the AdS boundary. In this case, the emitted Hawking radiation will be larger than the Hawking radiation coming back into the black hole since some radiation has escaped outside the boundary. The black hole, then, slowly evaporates. The more radiation we permit to escape from AdS, the faster the black hole evaporates. This procedure can be made precise within the context of AdS/CFT by coupling the boundary CFT to an auxiliary reservoir $\mathcal{H}_{rad}$, using absorbing boundary conditions. Furthermore, we assume that $\mathcal{H}_{rad}$ is a large holographic system, which allows the (holographic) encoding of the Hawking

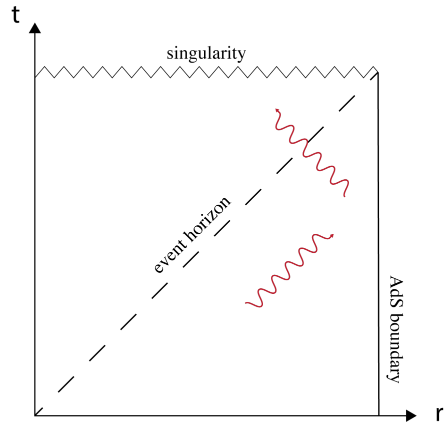


Figure 5.2: The Penrose diagram for a black hole with reflecting boundary conditions. Observe that the radiation (wiggly red lines) bounces on the boundary and returns to the black hole.

radiation into $\mathcal{H}_{rad}$ (see Figure 5.3). How can we incorporate H_{rad}, and thus the es-

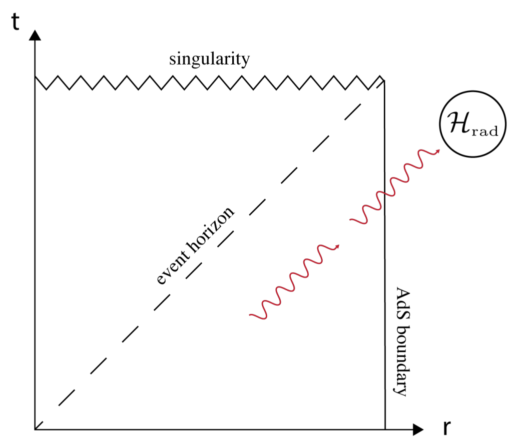


Figure 5.3: The Penrose diagram for a black hole with absorbing boundary conditions. Observe that some of the radiation (wiggly red lines) escapes from the boundary and is holographically encoded in $\mathcal{H}_{rad}$.

caping Hawking radiation, into the analysis of §5.1.3? Take, for instance, a quantum of late Hawking radiation L, which, after the Page time, is entangled with some interior modes B and the early Hawking radiation E, thus having a monogamy problem. As in the two-sided black hole case of §5.1.3, where the interior mode A was encoded in the left CFT, the interior mode B is encoded in the boundary theory. In particular, for one-sided evaporating black holes, the mode B is encoded in $\mathcal{H}_{rad}$.[18] Before the

[18]This statement is the core of the analysis of Penington (2019). Indeed Penington (2019) proves this claim using entanglement wedge reconstruction.

Page time t_p, $\mathcal{H}_{rad}$ encodes only the early Hawking radiation that escapes from the boundary of AdS. After the Page time, t_p, $\mathcal{H}_{rad}$ also encodes B, despite it being in the interior and thus unable to reach the boundary. This fact signals the breakdown of the semiclassical picture since we have two spacelike separated systems, B and E, which live in the same CFT $\mathcal{H}_{rad}$. Thus, there is no problem for the late Hawking radiation L to be entangled with both the interior mode B and $\mathcal{H}_{rad}$ (which encodes the escaped early radiation E), since, as in the case of the eternal AdS black hole, the first statement implies the second one. Since the interior mode B is encoded in $\mathcal{H}_{rad}$, entanglement with B implies entanglement with a mode in $\mathcal{H}_{rad}$. Equivalently, **(SD)** is violated since we have two spacelike separated systems, one inside and one outside the black hole, B and $\mathcal{H}_{rad}$ (which encodes E), which are nonetheless not distinct systems. Their not being distinct is a consequence of the fact that B is encoded in $\mathcal{H}_{rad}$, which equivalently means that B is a part of $\mathcal{H}_{rad}$. Thus, the two systems cannot be distinct. From the perspective of causal structures, we can analyse this situation in the same way in which we have studied the eternal AdS black hole of §5.1.3. The paradox is that it seems to be the case that $R_{ME}(L, B)$ and $R_{ME}(L, E)$, while $\neg R_{LC}(E, B)$, which by **(SD)** implies that B and E are two distinct systems, violating monogamy. However, as we have seen in this section, although E and B are spacelike separated, they are not distinct. Indeed, after t_p, B is holographically encoded in $\mathcal{H}_{rad}$ and therefore connected with E. This new connection can be encoded via R_{WH} and corresponds to a connection which is captured neither by R_{LC} nor by R_{ME}. Again, the generalised causal structure (defined via R_{WH}) captures the structure of the black hole that goes beyond the semiclassical approximation, which relies only on R_{LC} and R_{ME}. In particular, R_{WH} encodes those connections which show that the two systems B and E which, from the perspective of the causal structure of spacetime and entanglement, are distinct and separated, are interdependent and connected, thus violating **(SD)**.

Furthermore, this is again the same intuition of ER=EPR, that the semiclassical picture of spacetime crucially fails in the context of black holes in taking into account non-local connections that are neither causal curves nor entanglement relations. One of the main advantages of the causal structure approach is that we can naturally show the underlying strategy behind the different proposals for resolving the AMPS paradox we have studied thus far.

5.1.5 Conclusions

In this section, we have seen how the AMPS paradox appears to threaten the consistency of black hole physics and how dropping the implicit assumption of **spacetime distinctness** allows us to overcome it. The core of our work is the notion of causal structure. In particular, this notion helped us in highlighting the role of **(SD)** and in making explicit the link between various strategies for constructing the interior of the black hole, as explained in §5.1.3 and §5.1.4. Furthermore, causal structures make clear the sense in which ER=EPR-like connections imply a violation of semiclassical locality. Interestingly enough, **(SD)** being violated seems reminiscent of the basic idea of black hole complementarity Susskind et al. (1993), i.e. the idea that the interior and the exterior are two non-commuting descriptions of the same physics. It would then seem that the models we have analysed provide a precise incarnation of black hole complementarity by instantiating the violation of **(SD)**. However, understanding the connection between **(SD)** and black hole complementarity and how the ideas of this section can help black hole complementarity out of the disrepute in which some philosophers of physics hold it is beyond the scope of this section and will be left for future works.
Furthermore, it is far from clear that the non-locality involved in the constructions we have studied is not problematic. A task of extreme importance is understanding whether this non-locality leaks in the low energy regime Marolf and Polchinski (2016). Moreover, it is also essential to understand what it means for *locality* to be emergent. Most importantly, we have limited ourselves in this work to the AMPS paradox. However, AMPS is not the only paradox involved in constructing the interior of the black hole, as discussed in chapter §3. Most notably, in the context of AdS/CFT, we have not discussed the AMPSS paradox Almheiri et al. (2013) and related problems in the program of holographic interior reconstruction Harlow (2018). This will be the topic of the following section.

5.2 AMPSS

Before starting the analysis of the AMPSS paradox, and since we have used many different names for the various black hole paradoxes, let me recall our terminology. By firewall paradox, we mean a family of arguments allegedly showing the incompatibility of the effective field theory of gravity[19] and unitarity in the black hole interior. The first argument of this kind was *the AMPS paradox* of Almheiri et al. (2013), arguing that a unitary and semiclassical description of the black hole interior is incompatible with the monogamy of entanglement. In this context, a most fundamental question, which arguably was at the heart of Hawking (1976)'s worry is whether or not, and in which way, information[20] stored inside of the black hole can escape outside. In the context of AdS/CFT (Maldacena, 1999c; Witten, 1998),[21] where some of these questions can be made precise, Almheiri et al. (2013) argued that the question of interior reconstruction is trivial since the interior is *non-existent*. This conclusion is motivated by the observation that, in AdS/CFT, representing the interior degrees of freedom in the boundary CFT seems to lead to inconsistency. This argument is known as the *AMPSS paradox*. Note that the reasoning of AMPSS can be seen as a continuation of the arguments developed in Almheiri et al. (2013), and, in particular, as a critique of solutions to the AMPS paradox relying on what we will call *route (iii)*. Since the arguments presented in AMPS and AMPSS are formally different, we will use the following terminology:

- *AMPSS paradox* or *creation operator problem* (Harlow, 2016, p. 101): an argument against the existence of a holographic dual to creation operators in the black hole interior (§3.4).
- *AMPS paradox* or *monogamy problem*: an argument against the existence of a consistent description of the black hole interior based on the violation of the monogamy of entanglement (§3.3).
- *Firewall paradox*: any argument purporting to show the incompatibility of unitarity and effective field theory in the black hole interior.

We focus on the AMPSS paradox because it is relevant to the task of interior reconstruction since it constitutes an obstruction to defining an algebra of operators for the black hole interior in the CFT. Moreover, it allows us to introduce and explain some of the features of *state-dependence*, one of the most counterintuitive features of recent

[19]With effective field theory of gravity, we mean gravity as an ordinary effective quantum field theory, valid up to an energy scale Λ. This expression is an umbrella term for the three possible ways to have a working theory of quantum gravity without a fundamental one, cited in section §2.1.1.

[20]For ease of exposition, we generically speak of information, as customary in the physics literature, and as we have done for all this thesis.

[21]For a review on AdS/CFT see Ammon and Erdmenger (2015). For a philosophically-minded introduction, see De Haro et al. (2016). For an introduction to the ontology of dualities, which we do not discuss in this thesis, see Le Bihan and Read (2018).

attempts at representing quantum black holes.

In this part of the thesis, we mainly do two things:

- We highlight an implicit assumption of the AMPSS paradox, *semiclassical exactness* **(SE)**, i.e. that quantities defined at small coupling in perturbation theory are genuine physical quantities. We also argue that state-dependent interior reconstructions such as those of Papadodimas and Raju (2016), Almheiri (2018), and Penington (2020) avoid the paradox by violating **(SE)**. This violation arises because quantities defined in perturbation theory are associated with state-dependent operators, i.e. operators whose definition depends on the microstate of the quantum system under study, in this case, a black hole. While somewhat similar ideas have been floating around in the physics literature, this is the first time their basic conceptual structure is analysed in this manner, and that **(SE)** and its crucial role in the paradox is identified.

- We then argue that the violation of **(SE)** in state-dependent reconstruction follows from certain non-local connections. These non-local connections, whose instances include models of the ER=EPR conjecture, are associated with the violation of a further principle, *spacetime distinctness* **(SD)**, whose violation is crucial for the resolution of the AMPS paradox. We call the conceptual strategy behind these constructions *route (iii)*. The argument presented in §5.2.4, highlighting the connection between **(SD)** and **(SE)** or, more generally, ER=EPR-like non-local connections and failure of perturbation theory, and their subsumption into the larger conceptual framework of *route (iii)*, has never appeared in the literature.[22]

The violation of **(SE)** has far-reaching consequences on our understanding of physics even beyond quantum black holes. For example, it implies that perturbation theory, one of the Standard Model of particle physics' cornerstones, is not a reliable tool in Quantum Gravity (QG). Moreover, it also implies that the degrees of freedom described by perturbation theory are not physical quantities, not even approximately.

Note that we work in the context of the AdS/CFT correspondence, where interior reconstruction and related problems are stated most precisely. For the reader who is optimistic regarding the prospects of extending AdS/CFT techniques to more general spacetimes, our arguments show how and at which cost information can escape from

[22] See Cinti and Sanchioni (2021b) for more details on this point.

black holes. For a reader less optimistic regarding the prospects of AdS/CFT, we still provide some conceptual insights on the structure of the holographic representation of black holes and the conceptual strategy behind interior reconstruction. Given that the physics on which we rely for our analysis can be derived directly from the gravitational path integral (Almheiri et al., 2020b; Penington et al., 2022), without holographic input, this task is relevant even if one does not believe in AdS/CFT. However, we will work within the holographic context for the remainder of this section since it makes the argument's basic structure most transparent.

The section is structured as follows. In §5.2.1, we provide an in-depth analysis of the violation of **(SD)**. In §5.2.2, we review the AMPSS paradox and highlight the role of **(SE)**. In §5.2.3, we show how state-dependence avoids the AMPSS paradox by violating **(SE)** and how this violation follows from the basic structure of route (iii). In §5.2.4, we conclude.

5.2.1 More on the Violation of (SD)

In section §5.1.1 have seen that there are three ways to overcome the AMPS paradox:

route (i) Drop $R_{ME}(L, E)$, echoing the idea of Hawking (1976), where there is no entanglement between E and L since the Hawking radiation is completely thermal. The drawback of this proposal is that Hawking radiation being thermal entails that the physics of an evaporating black hole is non-unitary.

route (ii) Drop $R_{ME}(L, B)$, the proposal of Almheiri et al. (2013), where there is no entanglement between B and L. This proposal implies that the event horizon is not smooth. Instead, a firewall at the horizon separates the interior and the exterior. The drawback of this proposal is that the equivalence principle of GR does not hold.

route (iii) Drop **(SD)**. This third possibility consists of treating E and B not as distinct systems but as one quantum system. In this way, L is not entangled with two quantum systems but with just one.[23]

[23] Let us briefly comment on how identifying E and B corresponds to violating **(SD)**. An instance of the strategy described here is what in the physics literature is known as $A = R_B$ (Harlow, 2016, p. 108), with E and B playing the role of A and R_B. Take A and R_B as spacelike separated systems, each with an associated non-commutative algebra of operators. If $A = R_B$, then it follows that their operator algebras, too, are equivalent. However, since the algebras are non-commutative, an arbitrary operator associated with R_B will not commute with an arbitrary operator associated with A. Thus

The drawback of the latter proposal is that the causal structure of spacetime at the quantum level will be different from relativistic spacetime. Indeed, there will be robust counterfactual connections between L and E, not present in the spacetime and entanglement causal structures. This fact follows because they are encoded neither by R_{LC} or R_{ME} since B and E are spacelike separated, and there is no entanglement between them. Rather, these new connections are part of a new QG-specific causal structure which we describe in the remainder of this section.[24]

Let us briefly expand on the relation between *route (iii)* and ER=EPR. A good starting point is the eternal black hole in AdS (§5.1.3). This black hole is equivalently described as two horizons connected by a wormhole or as two entangled black holes (Maldacena, 2003; Maldacena and Maoz, 2004). As such, it is the simplest possible example of the ER=EPR conjecture, as remarked in Maldacena and Susskind (2013). As shown in Maldacena and Susskind (2013), the presence of the wormhole is equivalent to the failure of commutativity between a mode just inside the right horizon and a mode just outside the left horizon. This failure of commutativity between two spacelike separated quantum systems is a violation of **(SD)**. More general examples, such as an evaporating black hole, do not admit a straightforward analysis in terms of a wormhole described by semiclassical geometry (§5.1.4).[25] Nonetheless, the basic structure of *route (iii)* and the violation of **(SD)** is present also in these cases (Penington, 2020; Cinti and Sanchioni, 2021b). As such, *route (iii)* gives a generalisation of the basic strategy behind the ER=EPR resolution of the AMPS paradox,[26] and explains the conceptual strategy behind ER=EPR, showing that it also applies to different incarnations of the firewall paradox, such as AMPSS.

The debate between *route (i), (ii)*, and *(iii)* is still very much open, with the basic

we have spacelike separated systems whose algebras of operator do not commute, hence a violation of **(SD)**.

[24] This reasoning also helps to understand Maldacena and Susskind (2013)'s claim that in ER=EPR a measurement on E *disrupts* B. The crucial point is that the sense in which acting on E disrupts B is encoded in the fact that operators associated with E do not commute with operators associated with B, as Maldacena and Susskind (2013) themselves show. Thus, E and B do not correspond to distinct degrees of freedom in the microscopic theory given by the boundary CFT. Equivalently, acting on E means acting on B.

[25] By semiclassical geometry here we mean a geometrical picture such that one of the eternal AdS black hole, in which the failure of commutativity between two space-like separated quantum systems can be understood in terms of a geometrical wormhole stretching between the two horizons. We call it semiclassical since it arises already in the semiclassical limit without having to deal with a more fundamental quantum gravitational theory, as already discussed in §5.1.3.

[26] The validity of ER=EPR in these examples is still a conjecture. In particular, it would require purely quantum wormholes, which Susskind (2016a) calls *planckian wormholes* instead of the semiclassical construction employed for eternal AdS black holes.

approaches being still very much an object of discussion. Indeed, approaches along the lines of *route (i)*, i.e. involving non-unitarity, have been defended, among others, by Unruh and Wald (1995a), Unruh and Wald (2017), Penrose (2007, pp. 840-841), and Maudlin (2017b). Approaches in the spirit of *route (ii)*, besides the various firewall proposals (Marolf and Polchinski, 2013; Almheiri et al., 2013), could include those coming from the *fuzzball* program (Bena et al., 2012; Gibbons and Warner, 2013; Mathur and Turton, 2014a,b),[27] which aims to substitute the black hole with a peculiar string-theoretic construction in which at the horizon one encounters a cluster of strings.

For this section, however, we will focus on *route (iii)*. We do this for two reasons. First of all, as we will see in the following, *route (iii)* raises a host of profound philosophical questions regarding the status of perturbation theory and locality in QG. For this reason, we think that philosophical analysis of its underpinnings is crucial. Indeed, given the large amount of attention that solutions falling under the umbrella of *route (iii)* have received in the physical literature (Maldacena and Susskind, 2013; Penington, 2020; Almheiri et al., 2019b, 2021, 2020b), we believe that the time for a conceptual analysis of *route (iii)* is ripe. Second of all, route (i) and route (ii) both imply the violation of well-established principles of physics. Route (i) entails the violation of the equivalence principle of GR since it implies that freely falling observers do not experience the gravitational vacuum at the horizon but instead meet the firewall. Since the equivalence principle is one of the fundamental insights of GR, one should be careful about renouncing it. Route (ii) entails a violation of unitarity. However, the non-unitarity in Hawking's calculation comes from Plank scale effects, which occur at the end of the evaporation process. On the other hand, the AMPS paradox occurs long before this time, when semiclassical gravity and effective field theory are still approximately valid descriptions. As such, it is unclear how this violation of unitarity might arise in this context. Moreover, as we have observed at the end of §3.3, unitarity is a core aspect of quantum theory. It is to be retained in quantum gravity, at least insofar as one can trust arguments based on the AdS/CFT correspondence. As such, both (i) and (ii) seem to imply the violation of well-justified, though not immune from revision, principles. On the other hand, (iii) would mean, as we remarked above, that

[27] Let us mention that while fuzzballs seem to fit naturally with firewalls, it is nonetheless an interesting and important question whether or not a solution of the firewall paradox based on fuzzballs ultimately is an instance of *route (ii)*. Indeed, proposals such as the *fuzzball complementarity* of Mathur and Turton (2014b) seem to posit the validity of $R_{ME}(L, B)$ contra *route (ii)*. We do, however, leave this issue for future work, and, for the remainder of this article, take firewalls as the paradigmatic example of *route (ii)*.

the relativistic notion of locality does not apply to quantum spacetime. However, the justification for thinking this seems to be shakier than the other principles discussed. To our knowledge, there is no explicit argument in defence of this claim. As such, we take it that before accepting (i) or (ii), one should at least test the viability of (iii), as it requires milder revisions to the fundamental principles of physics than its alternatives. In principle, one should not expect *unitarity* or *the equivalence principle* to hold at any energy scale. However, *route (i)* and *(ii)* rely explicitly on the semiclassical picture rather than a non-perturbative QG computation. At the semiclassical level, it seems reasonable to expect *unitarity* or *the equivalence principle* to hold since, in this regime, we can describe physics using a general relativistic spacetime and quantum field theory propagating on it. Indeed, the firewall paradox is surprising because it allegedly shows the incompatibility of these principles in the semiclassical regime where we expected them to hold. Hence, for *route (i)* and *(ii)* to work, there must be significant deviations from known physics already in the semiclassical regime where QG effects, which should lead to these violations, are small. In other words, to save the validity of the semiclassical approximation, *route (i)* and *(ii)* must sacrifice either *unitarity* or *the equivalence principle* even in the absence of strong QG effects. On the other hand, *route (iii)* implies that the causal structure of relativistic spacetime does not hold at the quantum level. However, it seems reasonable to expect that spacetime should have quantum features in QG, and thus its causal structure should not be purely classical. Indeed, contrary to *route (i)* and *(ii)*, these deviations from the causal structure of relativistic spacetime are due to non-perturbative effects, as we are going to see in §6.[28] Hence, rather than abandoning *unitarity* or *the equivalence principle*, *route (iii)* abandons the semiclassical approximation since semiclassical gravity does not describe non-perturbative effects. Including such non-perturbative effects extends and appropriately corrects semiclassical

[28]Here, when we speak of non-perturbative effects, we are using theoretical physicists jargon. With non-perturbative effects, we do not mean effects that are *analytic*, i.e. that do not need perturbation theory. When one uses perturbation theory to compute a physical quantity, one can encounter situations where there are instantons field configurations (such as in Yang-Mills theory) or branes (such as in string theory). The expression one typically finds is of the form

$$Z(g) = \sum_n a_n g^n + e^{-A/g} \sum_n a_n^{(1)} g^n \ , \tag{5.5}$$

where g is the coupling constant of the theory. The first term in the sum represents the ordinary "Feynman" perturbation series, while the second term proportional to $e^{-A/g}$ is the instanton contribution. Since all derivatives of the function $e^{-A/g}$ vanishes if $g \to 0$, the Taylor series of the second term do not appear in perturbation theory, i.e. when g is small. Thus, the instanton contribution $e^{-A/g}$ is a non-perturbative contribution because it is invisible in the perturbative expansion.

gravity computations to the non-perturbative QG regime. Thus, the deviations from semiclassical gravity of *route (iii)* take place beyond the purely semiclassical regime, as one would expect. In contrast, for *route (i)* and *(ii)*, the semiclassical approximation should hold, and thus *unitarity* or *the equivalence principle* need to be abandoned. For these reasons, *route (iii)*, with its accompanying violation of **(SD)**, seems a promising route to construct a consistent physics of quantum black holes and thus deserving of conceptual and philosophical attention. Moreover, one can already find concrete instances of this approach in the physical literature, especially concerning the ER=EPR conjecture (Maldacena and Susskind, 2013; Papadodimas and Raju, 2013; Penington, 2020). However, we do not wish to claim that *route (iii)* is the only way out of the firewall paradox and that *route (i)* and *(ii)* should be discontinued. Indeed, which solution to the firewall paradox is the correct one is still an open problem requiring significant theoretical and experimental work. Rather, we wish to claim that *route (iii)* is an important competitor in the race to solve the firewall paradox and that the conceptual questions it raises on black holes and QG are worthy of philosophical attention.[29]

Within route (iii), the violation of **(SD)**, as we have seen in §5.1.3 and §5.1.4, follows from the presence of a new quantum type of connection, which modifies the causal structure of spacetime. These connections are captured by what we have called *generalised causal structure*, defined by the relation R_{WH}[30] of *being non-trivially connected*, where non-trivial connectedness obtains whenever there is a robust counterfactual correlation between two systems A and B. In the case of the evaporating black hole of Figure 3.5, route (iii) says that there is a non-trivial connection between E and B, which neither R_{LC} nor R_{ME} capture, thus implying that the interior and the exterior cannot be treated as distinct systems.[31] R_{WH}, in this context, is intended as a more general relation than R_{LC} and R_{ME}, which encompasses them both while also taking into account connections such as those between B and E which can be understood in terms of neither R_{LC} nor R_{ME} (but are rather of a more QG nature).[32] The connection

[29] For a detailed analysis of these models and for a defence of the claim that they provide instances of route (iii), see Cinti and Sanchioni (2021a).

[30] The subscript WH in R_{WH} stands for *wormhole*, referencing the ER=EPR conjecture.

[31] Since this connection makes their algebras of observables not mutually commuting.

[32] Just as an example of the kind of connectivity that might be involved, take the eternal AdS black hole, studied in Maldacena and Susskind (2013), where the QG connections encoded in R_{WH} can be given a semiclassical interpretation in terms of the wormhole connecting the two asymptotic regions. However, while we can give a semiclassical description of these connections in this idealised example, this fact does not hold for more general cases where this description is unavailable. Only a purely QG one makes sense (quite intuitively, Susskind (2016a) calls these *planckian wormholes*). Recall, as we mentioned above, that the relation between these wormholes, R_{WH}, and **(SD)** is crucial to understand

between R_{LC} and R_{ME} on one side and R_{WH} on the other is that R_{LC} and R_{ME} are embedded in the causal structure defined by R_{WH}. However, this embedding does not exhaust the connections encoded by the generalised causal structure, as is the case for the R_{WH} connection between B and E. In other words, the interior and the exterior of an evaporating black hole are non-trivially connected by R_{WH}, which, however, does not imply any observable violation of locality since we are dealing with non-perturbative contributions. Thus R_{WH} should not influence perturbative physics.

To see that these connections are non-perturbative, note that the original Hawking computation is exact to all orders in perturbation theory (Mathur, 2009; Almheiri et al., 2021). However, the connections encoded by R_{WH} underlie a correction to the original Hawking computation to ensure unitarity. Hence, they cannot be perturbative but need to be non-perturbative. More intuitively, Hawking's computation is perturbatively exact, and its perturbative background is a spacetime without any wormholes or other non-local, R_{WH}-like connections. Thus, these connections must be non-perturbative.[33]

5.2.2 Behind AMPSS

In section §3.4, we have seen how the AMPSS paradox threatens the possibility of reconstructing the interior of the black hole from $\mathcal{H}_{\text{rad}}$, and thus how it makes recovering the information in the interior impossible. In particular, the paradox concerns the existence of a way to map an interior creation operator $b^\dagger$ in the bulk semiclassical theory to a linear operator (which we also call $b^\dagger$) in the boundary CFT.

Let us now, however, highlight the crucial role played by a seemingly innocent assumption behind AMPSS' reasoning: the assumption that the interior creation operator $b^\dagger$ is a linear operator in $\mathcal{H}_{\text{rad}}$. Indeed, without such an assumption, the paradox cannot be derived. Suppose $b^\dagger$ is a non-linear operator.[34] Then (3.11) and (3.12) cannot hold as operator equations on the entire CFT Hilbert space since they explicitly assume linearity, and AMPSS' argument does not go through. (3.11) and (3.12)'s assumptions of linearity is equivalent to the standard commutation relations $[b, b^\dagger] = 1$ and $[b, b] = [b^\dagger, b^\dagger] = 0$ (written in natural units) holding. To see this point, note that

the connection between *route (iii)* and ER=EPR.

[33]Indeed Penington (2020) explicitly shows that the corrections associated to R_{WH} are non-perturbative.

[34]The foundations of non-linear operators are still very much in development. For recent progress in the basic physical and mathematical structure of non-linear operators, see Akers et al. (2022). A detailed exploration of the conceptual and philosophical foundation of non-linear operators goes beyond the scope of this thesis and is left for future works.

(3.11), since it involves a commutator, depends on the standard commutation relations. Moreover, moving from $b^\dagger b = 1$, i.e. the abstract definition of the left inverse, to (3.12) requires that $[b, b^\dagger] = 1$, as one can readily check with some simple computations. Lastly, dealing with linear operators and the standard commutation relations holding is the same since $[b, b^\dagger] = 1$ holds iff the commutation operation is closed under linear combinations, i.e. if the operators involved are linear.[35]

The assumption of linearity might seem quite natural since $b^\dagger$ is supposed to encode a genuine physical quantity in the bulk, and genuine physical quantities are linear operators in QFT. However, in the present holographic context, it is important to remember that the map giving the connection between $b^\dagger$ in the bulk spacetime and $b^\dagger$ as an operator in the boundary CFT encoding non-perturbative physics is an error-correcting code (Almheiri et al., 2015; Bain, 2020b). Moreover, such a map is known only in perturbation theory around $G_N = 0$, which means that we only know how to map CFT quantities to bulk quantities in the low curvature regime of the bulk spacetime.[36] Equivalently, the map is known only in the large N limit of the CFT. Then, to assume that $b^\dagger$ is a linear operator in the CFT is to assume that we can ignore any non-perturbative corrections to this map, which otherwise might not allow mapping $b^\dagger$ in the bulk to a linear operator in the CFT. Equivalently, we do not need to worry about any errors in this error-correcting code. Thus, we are assuming that, as long as we remain in a regime where perturbation theory is reliable, we can safely count an operator defined in perturbation theory ($b^\dagger$) as representing a genuine physical quantity. Equivalently, we can regard it as a linear operator even if we do not know its non-perturbative completion (encoded in the CFT). Moreover, since, in the present context, perturbation theory is around $G_N = 0$, we are defining operators coming from semiclassical gravity.[37] Thus, since $b^\dagger$ makes sense as a physical quantity in semiclassical gravity, it should be represented as a linear operator in the CFT. Let us call this assumption *semiclassical exactness*:

(SE): Quantities defined at small coupling in *perturbation theory* can be treated as *genuine physical quantities.*

[35] To see this point, consider the structure of creation and annihilation operators with regards to the harmonic oscillator, where analogous considerations about linearity apply since the operator algebra are analogous.

[36] For more on the implications of this point for interior reconstruction see Hayden and Penington (2019b).

[37] For this reason, in what follows, we use perturbative, semiclassical, and small coupling quite interchangeably.

Since genuine physical quantities for us are linear operators,[38] **(SE)** implies that quantities defined at small coupling in perturbation theory are linear operators. Thus, violation of **(SE)** leads, in this context, to a violation of linearity. Indeed, these are interrelated issues. Once we represent genuine physical quantities by linear operators, as standard in quantum theory, the violation of **(SE)** implies a violation of linearity. In the definition of **(SE)**, we focus more on the perturbation theory side because it clarifies its relation with state-dependence, in the sense that state-dependent operators are operators who are well-defined (linear) perturbatively but not non-perturbatively. Moreover, note that, in the present context, there is a difference between *calculated through* perturbation theory and *defined in* perturbation theory. While subtle, it amounts to this: quantities *defined in* perturbation theory are quantities required to satisfy certain conditions at lower order in perturbation theory, without the specific requirement that they approximate a linear operator non-perturbatively. On the other hand, when speaking of *calculating through* perturbation theory, it appears that one is implicitly assuming that such non-perturbative completion exists.

How reasonable is **(SE)**? First of all, let us note that, in general, non-perturbative corrections are tiny and can be ignored in semiclassical computations. Thus, when we try to understand how to represent bulk quantities in the CFT, it would seem reasonable to represent semiclassical quantities as linear operators as long as we do not try to carry out computations at strong coupling. While there would be non-perturbative corrections to these quantities, they would be too little to matter,[39] thus rendering **(SE)** valid. Indeed, failure of **(SE)** would imply a rather dramatic breakdown of perturbation theory, which would not be valid even for the small coupling regime where it should be reliable. One might even argue that the entire success of the Standard Model of particle physics, which famously relies on perturbative methods, speaks in favour of **(SE)**. Indeed, were **(SE)** to be violated, it would be unclear how the quantities typically computed in scattering calculations could be trusted. In this sense, to assume **(SE)** is to assume that we can trust perturbation theory in its regime of validity, an

[38] While certainly not exhaustive of the meaning of *genuine physical quantity*, we take this clause to be a necessary condition on the meaning of *genuine physical quantity* in a quantum theory.

[39] Note that non-perturbative corrections can be quite important physically. Think, for example, of quark confinement in QCD (Jaffe and Witten, 2006). However, for semiclassical quantities such as a creation operator in the interior of a black hole, non-perturbative corrections should not be crucial since they are of order $O(e^{-S_{BH}})$ and are thus exponentially suppressed by the large entropy of the black hole. Nonetheless, non-perturbative corrections do end up being relevant to the black hole interior, as we are going to see.

assumption on which most particle physics relies and which appears to be hardly contestable.

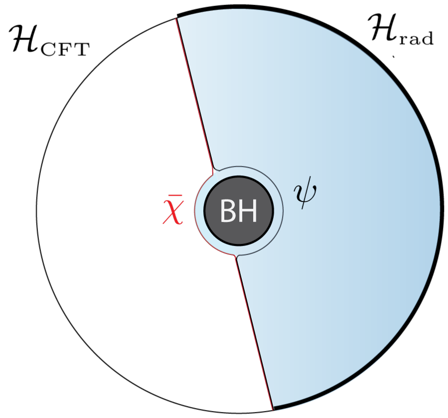
In the next section, however, we will see that extending **(SE)** from its usual reliable application in particle physics to the black hole interior is indeed problematic.

As we have seen in §3.4.1 and §3.4.2, to understand how bulk quantities are encoded in the CFT, we use entanglement wedge reconstruction. In particular, **(EWR)** states that a boundary region A encodes all the bulk physics in its associated entanglement wedge. In the following sections, our goal is to see how, when dealing with the black hole interior, **(SE)** is naturally violated within entanglement wedge reconstruction and how this fact implies a natural resolution of the AMPSS paradox (§5.2.3). Moreover, we are going to see how route (iii) naturally leads to the violation of **(SE)**, thus giving an overarching conceptual framework for black holes in QG (§5.2.4).

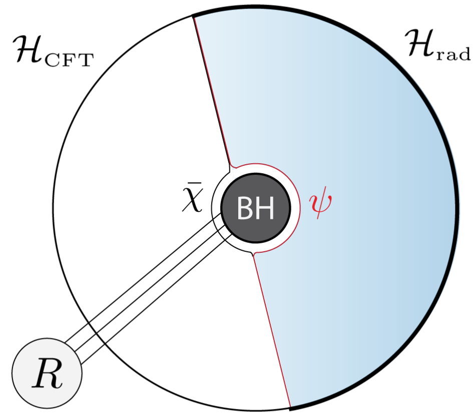
5.2.3 State-Dependence

When applying entanglement wedge reconstruction to an evaporating black hole to understand how the radiation can encode the interior, we are interested in understanding how subregions of the boundary CFT map to bulk quantities. In particular, we have seen in §3.4.1 that, given a region A in the CFT and its complement $\bar{A}$, if the state on $A \cup \bar{A}$ is pure, then the two regions have the same HRT surface. Moreover, one of them must necessarily include the interior in its entanglement wedge. Let A be $\mathcal{H}_{\text{rad}}$, and $\bar{A}$ be $\mathcal{H}_{\text{CFT}}$. Then we are certain we can always reconstruct the black hole interior from the boundary. In particular, as we saw in §3.4.2, after the Page time, the interior is encoded in $\mathcal{H}_{\text{rad}}$. However, if the state of $\mathcal{H}_{\text{rad}} \otimes \mathcal{H}_{\text{CFT}}$ is mixed, the HRT surfaces of the two subregions do not in general coincide. In particular, the black hole interior can be outside of the two entanglement wedges for sufficiently mixed states. To understand this claim, let us recall that the HRT surface is a quantum extremal surface which minimises S_{gen} and computes the entanglement entropy S_{EE} of a CFT subregion. S_{gen} is given by two terms, an area term $A(\chi)$ and a matter entropy term $S_{\text{bulk}}(\chi)$, as in the formula (3.7). Let us also assume that the Page time has passed, so that we expect the interior to be encoded in $\mathcal{H}_{\text{rad}}$ (§3.4.2).

In a black hole system after the Page time, there are two quantum extremal surfaces connected to $\mathcal{H}_{\text{rad}}$: a surface ψ, whose entanglement wedge includes just the exterior of the black hole, and the surface $\bar{\chi}$ of §4.3, whose entanglement wedge also includes



(a)



(b)

Figure 5.4: (5.4a) A schematic representation of a black hole (BH) whose interior is in a pure microstate, so that it lies within the entanglement wedge of $\mathcal{H}_{\text{rad}}$ (shaded) delimited by $\bar{\chi}$ (red). (5.4b) A schematic representation of a black hole (BH) whose interior is in a highly mixed microstate so that it does not lie within the entanglement wedge of $\mathcal{H}_{\text{rad}}$ (shaded) delimited by ψ (red). The mixed nature of the black hole microstate is obtained by entangling it with a reference system R. The reduced density matrix of the black hole will then be a mixed state. Note that R is not to be understood as an actual physical system in the world but rather as a tool that we use to aid in the representation of a black hole in a mixed state.

the interior of the black hole (see Figure 5.4). While for pure and lightly mixed states of the black hole, the HRT surface for $\mathcal{H}_{\text{rad}}$ is $\bar{\chi}$, for highly mixed states, the entropy $S_{\text{bulk}}(\bar{\chi})$ associated to fields in the interior of the black hole becomes so large[40] that the HRT surface switches to the surface ψ. ψ is the new HRT surface since $S_{\text{gen}}(\psi)$ is smaller than $S_{\text{gen}}(\bar{\chi})$. To see this point, first of all, let us assume that $A(\bar{\chi}) < A(\psi)$, and moreover note that $A(\psi) - A(\bar{\chi}) < A_{\text{hor}}$, as evident from Figure 5.4. For pure states, where the area contribution to S_{gen} dominates, we have that $\bar{\chi}$ is the correct HRT surface. However, take, for example, a state maximally mixed for the fields in the black hole interior, i.e. a mixture of all possible black hole microstates. This state is certainly highly mixed and gives a contribution to $S_{\text{bulk}}(\bar{\chi})$, and thus to $S_{\text{gen}}(\bar{\chi})$, of order S_{BH}, the black hole's microcanonical entropy. Since we know from black hole thermodynamics that $S_{\text{BH}} = A_{\text{hor}}/4G_N$, we have that, for such highly mixed states, the contribution S_{bulk} of the interior fields to $S_{\text{gen}}(\bar{\chi})$ is of order A_{hor}. Therefore, the interior

[40]Indeed, it becomes $O(1)$.

fields give a contribution exceeding $A(\psi) - A(\bar{\chi})$, which implies that $S_{\text{gen}}(\psi) < S_{\text{gen}}(\bar{\chi})$. For these highly mixed states, where highly mixed means that it gives a contribution to the entropy S_{gen} greater than $A(\psi) - A(\bar{\chi})$, ψ is the HRT surface, not $\bar{\chi}$. The upshot of this discussion is that the interior is no longer encoded in the radiation since the entanglement wedge of $\mathcal{H}_{\text{rad}}$ defined by ψ does not extend into the interior. Therefore, by **(EWR)**, we can no longer recover the information inside the black hole.

The switch between $\bar{\chi}$ and ψ is quite different from the phase transition between $\bar{\chi}$ and $\varnothing$ at the Page time described in §3.4.2. Indeed, the phase transition between $\bar{\chi}$ and $\varnothing$ happens in time and is thus a proper phase transition. On the other hand, the switch between $\bar{\chi}$ and ψ is not a function of time but of the mixedness of the boundary state. To understand how to resolve this difficulty, note that, as we mentioned before, the error-correcting map between bulk and boundary has non-perturbatively small errors. A specific pure bulk state is associated with a pure CFT state up to these non-perturbative errors in the map. However, when we consider mixed states, which are just combinations of pure states, these errors in the reconstruction map will add up. These errors become $O(1)$ for sufficiently mixed states and make the reconstruction map unreliable. The error is too large for error correction to work. In such a situation, the only way to have a reliable map is to limit its validity to a subspace of $\mathcal{H}_{\text{rad}}$ small enough to avoid highly mixed states, which would make error correction impossible. Equivalently, take a state which acts as a perturbative vacuum and construct the subspace of the whole Hilbert space comprising those states reachable by small perturbations of our perturbative vacuum. Perturbation theory makes sense within this subspace, and non-perturbative effects can be ignored.[41] This is the basic idea behind *state-dependent reconstructions of the interior*.[42] Therefore, a map between bulk and boundary exists only in a subspace of $\mathcal{H}_{\text{rad}}$ small enough to avoid the problems described above. If we were to pick a different subspace, we would have a different map. Thus, our maps depend on the state that we choose to define the subspace of $\mathcal{H}_{\text{rad}}$, hence the name state-dependence.

The upshot of this discussion is that we can still reconstruct the interior from $\mathcal{H}_{\text{rad}}$. However, this reconstruction is state-dependent, avoiding the highly mixed states that

[41]This perspective was the one originally adopted by Papadodimas and Raju (2013) before the introduction of quantum error correction methods in AdS/CFT.

[42]Originally introduced in Papadodimas and Raju (2013). Note that state-dependence, while being a proposed resolution to the AMPSS paradox and other issues in the research program of holographic black holes, and being supported by various computations, has not yet been rigorously proven (though see Akers and Penington (2022) for recent progress on this front).

cause problems. The language of error correction provides a natural framework to understand this state-dependence, which might otherwise seem mysterious. This observation clarifies the sense in which the mixed-state angle on state-dependence employed by Penington (2020)[43] and the entanglement angle employed by Harlow (2016)[44] can be seen as concrete incarnations of the two complementary perspectives given by **(SE)** and non-linearity on the issue of state-dependence. In particular, the entanglement observable behind Harlow (2016)'s route to state-dependence is the von Neumann entropy of the reduced density matrix of a CFT subregion. However, $S(\rho)$ is also a measure of how mixed a state is, thus leading to Penington (2020)'s route to state-dependence. Thus, the two perspectives appear to be complementary. However, we focus here on Penington (2020)'s perspective since when considering systems with infinite degrees of freedom, such as in a continuum treatment of the boundary CFT, $S(\rho)$ diverges and is not a well-defined observable. However, the quantum error correction approach of Penington (2020) remains valid even in this limit (Kang and Kolchmeyer, 2021). Let us now see how these ideas allow us to solve the AMPSS paradox.

Recall that the problem with AMPSS was that an operator whose existence we expected from the perspective of semiclassical physics did not make sense as a linear operator in $\mathcal{H}_{\text{rad}}$. This fact was encoded in the assumption of **(SE)**, which is crucial to deriving the paradoxical conclusion of AMPSS. However, if we accept that interior operators are encoded in a state-dependent way in $\mathcal{H}_{\text{rad}}$, then we have a natural way out of this conundrum. In particular, state-dependent operators are not linear operators on $\mathcal{H}_{\text{rad}}$, and as such, they violate **(SE)** and evade the AMPSS paradox, giving a consistent way to encode interior physics in the CFT. The operator $b^\dagger$, at the heart of AMPSS, is defined as that operator which satisfies equations (3.11) and (3.12). Since AMPSS assume **(SE)**, this implies that there must be a single linear operator in the CFT satisfying these equations. However, if we accept the possibility of state-dependent reconstructions, then, given two different subspaces $\mathcal{H}_A$ and $\mathcal{H}_{A'}$ in $\mathcal{H}_{\text{rad}}$, we are going to have a different representation of $b^\dagger$ for each of these subspaces. Hence, we are going to have an operator $b_A^\dagger$ which satisfies the equations (3.11) and (3.12) in $\mathcal{H}_A$, and not outside, and a different operator $b_{A'}^\dagger$ which satisfies these equations in $\mathcal{H}_{A'}$, but not outside.

Overall, these considerations imply that $b^\dagger$ is not represented as a linear operator in $\mathcal{H}_{\text{rad}}$

[43] Where the non-linearity of state-dependence is because sufficiently mixed states lead to large non-perturbative corrections, which make error correction impossible.

[44] Where the non-linearity of state-dependence is seen as the product of attempting to represent spacetime using an entanglement observable which is non-linear.

but as a non-linear quantity. Moreover, there is no contradiction between $b^\dagger$ lowering energy, as per (3.11), and $b^\dagger$ being invertible, as per (3.12), since the operator $b^\dagger_A$ giving the state-dependent representation of $b^\dagger$ in $\mathcal{H}_A$, for example, lowers energy only within $\mathcal{H}_A$, but in general, will raise energy outside of $\mathcal{H}_A$. Hence, it is not a many-to-one map on $\mathcal{H}_{\text{rad}}$, which implies that there is no obstruction to its being invertible. This way, we restore consistency with effective field theory since the problem raised by AMPSS had to do with the possibility of having a creation operator which lowers energy for any state in $\mathcal{H}_{\text{rad}}$. However, no such operator is under discussion when considering state-dependent reconstructions since $b^\dagger$ is not a linear quantity. Note that this way of resolving the paradox rests crucially on the violation of **(SE)** since we explicitly deny that quantities which make sense in perturbation theory, like $b^\dagger$, need to be represented as linear operators. This observation avoids the pitfalls of AMPSS. If **(SE)** were not violated, then state-dependence would not have been possible. As such, the violation of **(SE)** is the crucial conceptual move underlying the state-dependent strategy for solving AMPSS.

Note that, by the previous discussion, failure of **(SE)** implies that perturbation theory is not reliable. However, this claim does not hold in full generality; indeed, there are many regimes in QG where perturbation theory is perfectly reliable (see Wallace (2022) for some examples). Rather, what the violation of **(SE)** implies is that in QG, contrary to the lore of standard QFT, even in regimes where a semiclassical perturbative picture *prima facie* makes sense, this perturbative picture, nonetheless, is not always reliable. As we have seen, an example of this phenomenon is the black hole interior. Another possible example, though, in this case, the theoretical framework is much more uncertain, is the holographic description of the early universe (Hartman et al., 2020; Chen et al., 2021; Balasubramanian et al., 2021). While its full evaluation goes beyond the scope of this section, a proposed criterion to neatly separate between regimes in QG where perturbation theory applies or not, based on considerations coming from computational complexity, has been proposed in Akers et al. (2022).

Nevertheless, how can we justify the violation of **(SE)**? Should we assume it to make interior reconstruction possible? In the next section, we are going to discuss how to motivate the violation of **(SE)** and the use of state-dependence from the perspective of route (iii).

5.2.4 From Route (iii) to State-Dependence

From the perspective of route (iii), in which **(SD)** is violated, the violation of **(SE)** is natural. First of all, however, let us explain the relation between the violation of **(SD)** and interior reconstruction. In §5.1.3 and §5.1.4, we argued that the violation of **(SD)** occurs due to the presence of robust counterfactual connections, represented by R_{WH}, between the interior of a black hole and its Hawking radiation. Since these connections lead to the algebras of observables of the interior and the Hawking radiation not commuting, in principle, an operator acting on the Hawking radiation should also act on the interior of the black hole since they are not independent degrees of freedom. Therefore, one can expect that we can reconstruct at least part of the interior of a black hole by taking some (possibly complicated) combination of operators acting on Hawking radiation.

Let us now see why R_{WH} leads to state-dependence, the violation of **(SE)**, and the avoidance of the AMPSS paradox through the following argument:

(1) $\mathbf{R_{WH} \to \neg(SD)}$

By the discussion in §5.2.1, R_{WH} encodes the existence of certain non-local connections, which appear when studying the interior of black holes in AdS/CFT. These connections lead to the violation of **(SD)**, i.e. spacetime distinctness.

(2) $\mathbf{R_{WH}(B, E) \to [E, B] \neq 0}$

For an exterior mode part E in the early Hawking radiation and its interior partner B, violation of **(SD)** means that: $R_{WH}(B, E) \to [E, B] \neq 0$.

(3) $\mathbf{[E, B] \neq 0 \to [b_i^\dagger, b_j^\dagger] \neq 0}$

Let $b_i^\dagger$ and $b_j^\dagger$ be schematic representations of creation operators for respectively B and E. Since the creation and annihilation operators are the generators for the operator algebras of B and E, non-commuting operators between B and E imply non-commuting generators. Hence, (2) implies that $[b_i^\dagger, b_j^\dagger] \neq 0$.

(4) $\mathbf{[b_i^\dagger, b_j^\dagger] \neq 0 \to}$ **non-linearity**

Hence, there cannot exist a linear creation operator associated with B, since by definition (linear) creation operators associated with different modes should satisfy $[b_i^\dagger, b_j^\dagger] = 0$, i.e. commute. Failure to do so implies that we are not dealing with creation operators or that those creation operators cannot be linear. Since, by assumption, $b_i^\dagger$ and $b_j^\dagger$ are creation operators, they cannot be linear. Thus, one gets a violation of linearity.

(5) $\mathbf{R_{WH}} \rightarrow \Big[[\mathbf{b_i^\dagger}, \mathbf{b_j^\dagger}] \approx \mathbf{0}$ **at small coupling**$\Big]$

However, the starting point of AMPSS is essentially correct: at small coupling, we can define a creation operator for B (and one for E). Thus, $[b_i^\dagger, b_j^\dagger] \approx 0$ at small coupling since the R_{WH} connections of (1) must be non-perturbative (§5.2.1). The R_{WH} connections must be non-perturbative to avoid massive departures from semiclassical spacetime already in the perturbative semiclassical regime. In other words, relativistic locality appears to be observed, i.e. we do not see violations of **(SD)** in the low energy physics around us. Thus, violations of **(SD)** must involve some non-perturbative effect going beyond a low energy approximation.[45] In other words, if **(SD)** were violated already at small coupling in perturbation theory, the perturbation theory itself would not make sense since the relevant perturbative background would not be suitable for this task.

(6) **non-linearity** & $[\mathbf{b_i^\dagger}, \mathbf{b_j^\dagger}] \approx \mathbf{0}$ **at small coupling** $\rightarrow$ **state-dependence**

From (3) and (5), we can see that the creation operator for an interior mode is well-defined only insofar as we ignore non-perturbative effects, i.e. this operator is state-dependent. To see why, recall, as we saw in §5.2.3, that we could define subspaces of the whole Hilbert where perturbation theory makes sense, and non-perturbative effects can be ignored. Within these subspaces, $[b_i^\dagger, b_j^\dagger] \approx 0$ while outside $[b_i^\dagger, b_j^\dagger] \neq 0$ since outside the non-perturbative effects which make the commutator different from zero become relevant. Thus, inside the subspace, $b^\dagger$ is well-defined as an operator while outside, it is not, making $b^\dagger$ a state-dependent quantity.

(7) **State-dependence** $\rightarrow \neg$**(SE)**

Since $b^\dagger$ is well-defined in perturbation theory but not non-perturbatively, we have a violation of **(SE)**, which is the mechanism by which we avoid the paradoxical conclusion of AMPSS.

(conclusion) $\mathbf{R_{WH}} \rightarrow \neg$**(SE)**

[45] Going back to the intuitive picture of the eternal AdS black hole, this would be equivalent to saying that there are semiclassical wormholes outside of the horizon, which goes against the definition of a wormhole in GR.

In conclusion, the existence of R_{WH} connections which violate **(SD)** quite naturally leads to state-dependence and the violation of **(SE)**.

Thus, route (iii), a strategy for resolving the AMPS paradox, doubles as a solution to the AMPSS paradox. Route (iii) and R_{WH} provide a unified conceptual strategy behind the resolution of the AMPS and AMPSS paradoxes and show that the same type of phenomena, i.e. non-perturbative corrections from QG, determine the emergence of a variety of non-classical behaviours. These behaviours range from non-local connections ($\neg$**(SD)**) to the breakdown of perturbation theory and the semiclassical approximation ($\neg$**(SE)**). Indeed, insofar as route (iii) can be motivated directly from the construction of explicit models in AdS/CFT, without resorting to speculative steps,[46] then this discussion suggests that route (iii) will be sufficient to have a consistent description of black hole physics without the need for further assumptions.

5.2.5 Conclusions

In this section, we have seen how the interplay between holographic interior reconstruction and the AMPSS paradox naturally leads to the notion of state-dependence. Moreover, we have also argued that such state-dependence is naturally understood as a consequence of route (iii), thus supporting the idea that there is a unified conceptual structure behind recent work on the physics of black holes beyond General Relativity. However, much remains to be done to understand the far-reaching consequences of the violation of **(SE)**, starting from its implications for the reliability of perturbation theory in QG and, even more importantly, for the physical status of perturbative approximations. Moreover, it is also essential to give a more intrinsic characterisation of route (iii) beyond the violation of **(SD)** and **(SE)**, in particular in light of recent results relying on the gravitational path integral (Almheiri et al., 2020b; Penington et al., 2022). These issues will be topics for future work.

[46] Indeed, recent calculations suggest that even AdS/CFT might not be necessary and that only the semiclassical gravitational path integral is required (Almheiri et al., 2020b; Penington et al., 2022). Nevertheless, the extension of these results beyond AdS/CFT remains speculative.

Chapter 6

A First Take on Black Hole Ontology in Quantum Gravity: No Membrane at the Black Hole Horizon?

This chapter aims to investigate the possible implications of one specific resolution of black hole paradoxes, namely *route (iii)*, for the ontology of quantum gravity and black holes. The resolution in question addresses the apparent inconsistencies between classical general relativity and quantum mechanics in the context of black holes. In particular, the focus of this chapter is the analysis of the structure of the black hole event horizon and its fundamental status. The analysis will likely explore the implications of route (iii) resolution on the event horizon. Overall, this chapter aims to contribute to a deeper understanding of the nature of black holes and quantum gravity.

The first steps in understanding metaphysical and philosophical aspects of black holes in quantum gravity have been made in Wallace (2018) and Wallace (2019): the first paper shows that black holes can be considered thermodynamical objects in the fullest sense, while the second paper analyses the statistical mechanical underpinning of black hole thermodynamics. Both papers touch upon the structure of the event horizon of the black hole. Wallace (2018) says that, in order to interpret black holes as thermodynamical objects in General Relativity, a phenomenological membrane, called the stretched horizon, should be posited just outside the actual black hole horizon. Wallace (2019) argues, moreover, that in order to have a statistical mechanical underpinning of black hole thermodynamics from the point of view of an exterior observer, a real membrane, made of black hole microstates, should be posited at the black hole horizon.[1]

The goal of this chapter is to continue the investigation of the black hole horizon in light of the *route (iii)* resolution of black hole paradoxes. In the previous chapter, we focused on the implications of the AdS/CFT correspondence and the entanglement wedge reconstruction program for this resolution. However, in this chapter, we aim to have an even more broad and general discussion of the black hole paradoxes. The material presented in chapter §5 concerning the *route (iii)* resolution of the firewall paradox was heavily dependent on AdS/CFT correspondence techniques, such as the entanglement wedge reconstruction program. As argued in the introduction of this thesis (§1.2), this is not a tremendous bug since AdS/CFT and holography can be considered, somehow, a progressive research program, and it is worth exploring its consequences. However, in the present chapter, we would like to give a broader discussion on black hole paradoxes, building on two works (Penington et al. (2020); Almheiri et al. (2020b)) that are considered within the high-energy physics community, crucial for resolving the firewall paradox. They can be intended as a generalisation of Penington (2020) to more general theories of gravity and provide a broader perspective on how to approach the resolution of black hole paradoxes. In particular, the two works we will be building on in this chapter, Penington et al. (2020) and Almheiri et al. (2020b), have been considered groundbreaking in the field of high-energy physics for their approach to resolving the firewall paradox. They provide a new perspective on how to understand the structure of black hole horizon and its relationship with quantum mechanics by using new techniques and ideas, such as the *replica trick* and the *island formula*, respectively. Through a detailed analysis of these works, we aim to provide a deeper understanding of the black hole paradoxes and the *route (iii)* resolution, and how it relates to the task of understanding the fundamental structure of the black hole horizon.

Penington et al. (2020) and Almheiri et al. (2020b) revisited Hawking (1975)'s approach to semiclassical gravitational physics, a crucial ingredient in the firewall paradox. By studying the gravitational path integral, i.e. considering gravity as an effective field theory (what we have called in §3.1.1 perturbative quantum gravity), Penington et al. (2020) and Almheiri et al. (2020b) showed that Hawking did not consider some small non-perturbative corrections which, despite being relatively small, drastically change

[1] By exterior observer, I mean an observer far away from the black hole. For the rest of this section, I will use exterior observer and asymptotic observer interchangeably.

the physics of black hole evaporation. Their results provide a unitary firewall-less picture of the black hole evaporation process. These non-perturbative corrections that appear in the gravitational path integral are responsible for the emergence of the generalised causal structure with its relation R_{WH}, which was discussed in the previous chapter.

Before beginning, it is important to address potential concerns about the applicability of the arguments and make clear the connection with the physics presented in the previous chapters. In order to describe the physics of black holes in quantum gravity, this chapter will make use of two assumptions: the *central dogma* and the *entanglement wedge reconstruction*. The central dogma has been formally derived in the context of string theory Strominger and Vafa (1996) and in the context of AdS/CFT correspondence in Maldacena (1999a). Similarly, the entanglement wedge reconstruction is a theorem within the AdS/CFT correspondence (Dong et al., 2016; Jafferis et al., 2016). Therefore, the results presented in this chapter, as well as those discussed in relation to the AMPSS paradox, are well-established within the framework of AdS/CFT correspondence. Additionally, it should be noted that the work of Wallace (2019) also presents his results within the context of the AdS/CFT correspondence, making this work an in-depth investigation of the black hole horizon within the same approach to quantum gravity. However, I have chosen to present the arguments in this chapter in a somewhat independent way from AdS/CFT, following the work of Penington et al. (2020) and Almheiri et al. (2020b). These works are considered significant as they studied the process of black hole evaporation starting from the gravitational path integral and without holography, meaning that their results may be extendable to non-holographic theories of quantum gravity. However, it should be noted that these calculations were performed using a simplified two-dimensional gravitational model known as *JT gravity* (Saad et al., 2019).[2]. While this may raise concerns about the generalizability of the results, it is understood within the string theory physics community that the explicit calculations performed in JT gravity do not undermine the overall structure and applicability of the arguments presented in the work of Penington et al. (2020) and Almheiri et al. (2020b). To conclude, the ontological investigation of black holes in quantum gravity that I am providing in the present chapter is well-grounded within AdS/CFT and could be extendible to non-holographic theories of quantum gravity if the results

[2]For a philosophical discussion on the use of two-dimensional gravity as a toy model for general relativity see Fletcher et al. (2018)

of Penington et al. (2020) and Almheiri et al. (2020b) turn out to be extendible beyond JT gravity.

The specific question I am going to address in this chapter is whether the membrane posited by Wallace (2019) is still necessary in order to have a statistical mechanical underpinning of black hole thermodynamics, given these recent theoretical results. The chapter is structured as follows: §6.1 clarifies the phenomenological status of the stretched horizon, while §6.2 is a reconstruction of Wallace (2019)'s argument for positing a quantum membrane at the event horizon of a black hole. §6.3 is a conceptually oriented presentation of the results of Penington et al. (2020), and Almheiri et al. (2020b), which provide a consistent picture of the black hole interior. §6.4 presents four arguments against the necessity and the consistency of having a quantum membrane at the horizon. §6.5 concludes.

6.1 The Stretched Horizon

Starting with Bekenstein (1973), and Hawking (1975), black holes have been considered thermal objects among physicists. Indeed, when an object is thrown into a stationary black hole, i.e. a black hole which is in a certain equilibrium state, the black hole evolves dynamically to a new equilibrium state, following[3]

$$\frac{\kappa}{8\pi G_N} dA_{\text{hor}} = dM - \Omega dJ \ , \tag{6.1}$$

where κ is the surface gravity, A_{hor} is the area of the event horizon, M is the mass of the black hole, and Ω and J are the angular velocity and the angular momentum of the black hole respectively. On the surface, this formula appears to be very similar to the second law of thermodynamics. Indeed, if we postulate that the black hole has an entropy S_{BH} and a temperature T, given by

$$S_{\text{BH}} = \frac{A_{\text{hor}}}{4G_N} \quad , \quad T = \frac{\kappa}{2\pi} \ , \tag{6.2}$$

[3]In this review, I do not consider certain complications which go beyond the scope of the section, of which I give a lightning-quick review in this footnote. If one considers also charged black holes, one can add a term like ΦdQ in the equation (6.1), where Φ and Q are respectively the chemical potential and the charge of the black hole. If one considers a black hole in a non-flat slowly dynamically evolving background, one can add a term like $\mathcal{T} dL$ (or PdV) in the equation (6.1), where $\mathcal{T}$ is the tension of the black hole with respect to the environment and L is the curvature of the background (while P and V are the pressure and the volume of the black hole). For a complete discussion of these additional terms in equation (6.1), and issues related to them, see Armas et al. (2016).

as Bekenstein (1973) and Hawking (1975) did, we find exactly the first law of black holes thermodynamics, i.e.

$$dM = T dS_{\rm BH} + \Omega dJ \ . \tag{6.3}$$

Even though (6.1) is just a formal rewriting of the Einstein equations, and the definitions given in (6.2) are only admitted at the quantum level, i.e. when considering quantum fields in the surroundings of the black hole, as proposed by Hawking (1975), interpreting (6.1) as the second law of thermodynamics also at the classical level is very tempting. However, there are some obstacles to do this. First, a black hole in general relativity is just a *place* in spacetime. Indeed, a standard definition of black holes that one receives from a physics textbook is something like "a region of spacetime where gravity is so strong that nothing can escape from it." However, it is not evident how a place in spacetime can have properties, change in time, be in equilibrium. Secondly, the black hole, as seen by an asymptotic observer, seems to be disconnected from its exterior. Indeed, the black hole horizon is a null hypersurface in spacetime. If one tries to throw an object into a black hole, from the point of view of an exterior observer, it will never reach the horizon, and therefore it will not perturb the black hole. From the perspective of an asymptotic observer, how can an object be thrown into a black hole, perturb it, and originate the dynamical evolution given by (6.1)?
In order to provide a reasonable explanation for these worries, Susskind et al. (1993) postulated the existence of a new surface, called the stretched horizon, defined as follows:

(SH) The **Stretched Horizon** is a time-like surface placed around the event horizon of the black hole at a Plank-length proper distance from it.[4]

With this definition, one can assign properties of the black hole to **(SH)**. From this point of view **(SH)** is a two dimensional (potentially charged) viscous fluid around the black hole, and in this way the interpretation of (6.1) as the second law of thermodynamics becomes sensible. Indeed, since **(SH)** is a standard fluid, it can have properties, can evolve in time, and since it is time-like an object thrown into the black hole in equilibrium can interact with it also from the point of view of an exterior observer. Thus, the second law of black hole thermodynamics follows accordingly, and **(SH)** describes the physics of a black hole as seen by an asymptotic observer.

[4] The distance between the stretched horizon and the true event horizon of the black hole is so small that no real particles can cross the stretched horizon and come back (only virtual particles can).

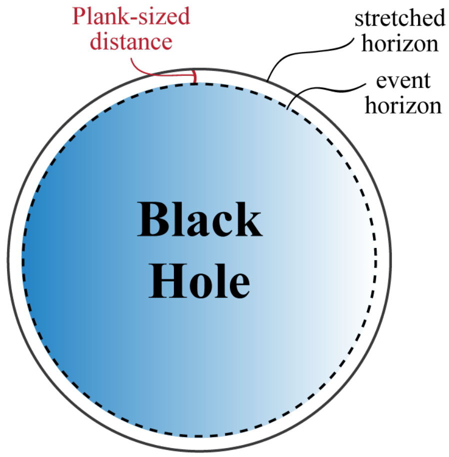


Figure 6.1: An illustrative picture of a black hole with its stretched horizon.

From the point of view presented above, it seems that **(SH)** is just a fictitious surface introduced to solve the puzzle created by our intuition that black holes should be classical thermodynamical objects. It seems, indeed, that we just adopted a mathematical trick to solve the problematic interpretation of equation (6.1). It is then natural to interpret **(SH)** as a *phenomenological* surface, useful for characterising the behaviour of a black hole in general relativity from the point of view of an asymptotic observer, but without any fundamental physical import.
Furthermore, the idea that **(SH)** should be just a phenomenological surface is also supported by consideration of the equivalence principle. Indeed, consider a Schwarzschild black hole in General Relativity described by the metric

$$ds^2 = \left(1 - \frac{r_s}{r}\right) dt^2 + \frac{dr^2}{1 - \frac{r_s}{r}} + r^2 d\Omega_2^2 \; . \tag{6.4}$$

This black hole solution, as it is well known, has an event horizon at $r = r_s$, where r_s is the Schwarzschild radius. The metric (6.4) describes the black hole as seen by an asymptotic observer. However, if we perform a change of coordinates suitable for describing the experience of an infalling observer, it is possible to study the structure of the spacetime in a neighbourhood of the event horizon. The "zoomed-in" black hole metric, which represents the near-horizon metric of a Schwarzschild black hole, is

$$ds^2 \approx -\rho^2 d\tau^2 + d\rho^2 \; , \tag{6.5}$$

which is conformally equivalent to ordinary flat Minkowski spacetime $ds^2 = -dt^2 + dx^2$.

Thus, within General Relativity, an observer who is free-falling into a black hole does not encounter anything special at the event horizon $r = r_s$. This observation is just an instance of the equivalence principle of General Relativity, which states that a free-falling observer does not feel the effect of gravity. Therefore, **(SH)** should be just a phenomenological surface which describes, at a non-fundamental level, the black hole as seen from an asymptotic observer, for otherwise the equivalence principle would be violated. If that were not the case when crossing the event horizon, one would encounter **(SH)** and interact with it, thus experiencing something different from free fall, contra the equivalence principle.

6.2 The Quantum Membrane

Considering black holes as thermodynamical objects is fully justified just when taking into account General Relativity coupled with Quantum Field Theory. In this framework, as shown reviewed in §3.1.1, black holes have a temperature given by (6.2), which fixes the proportionality constant between the entropy S_{BH} and the area A_{hor} to a value which gives the second law of thermodynamics, given by equation (6.3). Moreover, at the quantum level, black holes, as ordinary thermal objects, radiate and exchange heat with other surrounding objects. Indeed, as initially proposed in Hawking (1975), black holes in semiclassical gravity emit thermal radiation, called Hawking radiation. The fact that black holes radiate makes them thermally connected with objects in their exterior via Hawking radiation. Therefore, black holes seem to be thermodynamic objects in the fullest sense only when taking into account quantum effects.

One could infer, from what we have said so far, that, as for ordinary thermodynamic objects, there should be a statistical mechanical underpinning of black hole thermodynamics.[5] In other words, are there black hole microstates from which we can derive macroscopic thermodynamic-like properties? The proposal of Wallace (2019) for such statistical mechanical underpinning is to promote **(SH)** to a quantum membrane:

(QMP) **Quantum Membrane Paradigm**: With respect to any physical process taking place outside the stretched horizon of a stationary or near-stationary black hole, **(SH)** may be treated as a quantum-mechanical system at (or near) thermal equilibrium, with density of states given approximately by $N \approx \exp S_{\text{BH}}$.

[5] Though see Wuthrich (2017) for arguments against this line of thought.

In order to understand the necessity of **(QMP)**, we need to consider Hawking (1975)'s work on gravity as an effective field theory.

Gravity, as well-known, is a non-renormalisable theory.[6] Indeed, if we try to apply standard quantisation techniques to the gravitation field, we get an ill-defined theory. These issues arise if we want the quantisation of Einstein's theory to be well-defined at all energy scales. If we instead treat the quantisation of the gravitational field as an effective theory, i.e. a theory which is valid up to an energy cut-off scale Λ, there is no problem with renormalizability, as the loop momenta are cut-off. We can thus write a path integral for the gravitational field while keeping in mind that it gives a coherent description of physics only at scales below the cut-off scale Λ. Instead of the cut-off Λ, we can equivalently regard the cut-off black hole path integral as describing the black hole in a box of surface A_{box}. The gravitational path integral is given by the following expression:

$$\log Z(\beta) = \int_{\beta} \mathcal{D}g\mathcal{D}\phi \exp\{-S[g,\phi]\} \ . \tag{6.6}$$

Even though, to study physics at energies higher than Λ we would need a complete theory of quantum gravity, we can still draw some conclusions about the structure of black holes at the semiclassical level described by the cut-off path integral.

Our goal, following Hawking (1975), is to study quantum effects on a Schwarzschild black hole background. As for ordinary thermal objects, we need to derive macroscopic thermodynamic-like properties from the thermal partition function $Z = \text{Tr}\left[e^{-\beta H}\right]$. The thermal partition function of a quantum system in equilibrium is mathematically equivalent to the Euclidean time evolution of that quantum system on a circle of length β. In particular, by studying the path integral in Euclidian time on a circle of length β, we derive the thermodynamic properties of a system in thermal equilibrium at temperature $T = 1/\beta$. We thus consider the Euclidean path integral (6.6) on a Schwarzschild black hole background. The metric of a Schwarzschild black hole, represented in figure 6.2, in Euclidian-time coordinates, is

$$ds^2 = \left(1 - \frac{r_s}{r}\right) dt_E^2 + \frac{dr^2}{1 - \frac{r_s}{r}} + r^2 d\Omega_2^2 \dots \ , \qquad t_E = t_E + \beta \ , \tag{6.7}$$

[6] For a review of renormalisation in the context of QFT see Williams (2018).

where t_E is the Euclidean time (the Wick rotation of standard Lorentzian time, $t_E = it$), r_s is the Schwarzschild radius,[7] and the last equation in (6.7) expresses the fact that the Euclidean time should be periodic (which is a consequence of the fact that we study the Euclidean time evolution on a circle of length β). The metric (6.7) has the feature that it shrinks to zero at the horizon radius $r = r_s$. In order to preserve the equivalence principle of general relativity, we need the metric around $r = r_s$ to be smooth, which fixes β to be $\beta = 4\pi r_s$. It is this smoothness requirement that fixes the temperature of the black hole to be $T = \kappa/2\pi$, i.e. that of (6.2).

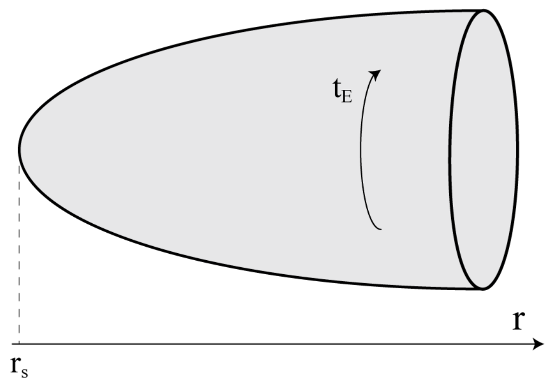


Figure 6.2: A Schwarzschild black hole in Euclidean time.

We just derived from the smoothness condition that the black hole has a temperature. We can thus consider the black hole to be an ordinary thermal object. If this is the case, then we can associate an entropy to the black hole given by $dS = dE/T = dM/T$, as we do for all other ordinary objects with a temperature.[8] Integrating this equation, and using the relations for r_s and β, we can derive the standard formula for black hole entropy:

$$S_{\text{BH}} = \int \frac{dM}{R} = \frac{A_{\text{hor}}}{4G_N} \ . \tag{6.8}$$

In conclusion, the requirement of the smooth horizon in the path integral formulation of the partition function for a Schwarzschild black hole fixes the temperature and the entropy to be the one of (6.2).

The question we may ask now is if the entropy we just introduced can be derived directly

[7] Which can be expressed in terms of the mass of the black hole as $r_s = 2GM/c^2$. This quantity is the only feature of the black hole present in the metric (6.7).

[8] This is a standard procedure in experimental physics when dealing with objects with a temperature.

from the path integral, i.e. from the partition function. If this is the case, we would have the statistical mechanical underpinning we desired. We thus turn to consider an expansion of the path integral (6.6) around the classical gravitational background given by (6.7), i.e. the Schwarzschild black hole:

$$\log Z(\beta) = -S_{\text{EH}}^{\text{classical}} + \log Z_{\text{quantum}} \ , \tag{6.9}$$

where $S_{\text{EH}}^{\text{classical}}$ is the Einstein-Hilbert action evaluated on the classical background (6.7), which is the gravitational part of the path integral, and Z_{quantum} is the quantum fields contribution, obtained by computing the partition function of the quantum fields on the fixed classical background (6.7). One can then compute the entropy of the black hole by:

$$S_{\text{BH}} = (1 - \beta\partial_\beta) \log Z(\beta) = \frac{A_{\text{hor}}}{4G_N} + S_{\text{matter}} \ , \tag{6.10}$$

using the saddle-point approximation.[9] S_{matter} represents the von-Neumann entropy of the quantum fields in the surroundings of the black hole.[10]

In order to better appreciate the difference between the entropy computed in (6.8) and the one computed in (6.10), let me first introduce two different concepts of entropy that are commonly used in discussions related to the black hole paradoxes:

S_{coarse} **Coarse-grained (thermodynamic) entropy**: Consider a system with a set of global properties A_i, which can be measured, and for which the microstates are unknown. In this case, there could be different density matrices describing the same global situation with properties A_i. The coarse-grained entropy is computed by maximising over all possible density matrices giving rise to the same global properties A_i, as

$$S_{\text{coarse}} = \max_{\tilde{\rho}} \left[\text{Tr} \left(\tilde{\rho} \log \tilde{\rho} \right) \right] \tag{6.11}$$

[9]As we are going to see in §6.3, the implementation of the saddle-point approximation is the main problem with this derivation.

[10]The S_{matter} term in the derivation of the black hole entropy from the path integral was already present in Bekenstein (1973)'s description. However, Bombelli et al. (1986) understood that also Hawking radiation should contribute to the matter entropy. Indeed, even if the entropy of the vacuum in Quantum Field Theory is zero, if one divides the space into parts, as the horizon of the black hole does, then the vacuum, when restricted in a region of space, say the exterior of the black hole, carries some entropy which is the entanglement entropy between the interior and the exterior of the black hole. Also, this entropy should contribute to S_{matter}, and the entropy of the Hawking radiation precisely gives this contribution.

The coarse-grained entropy is thermodynamic entropy, and it is not a precise measure. Since we do not know the precise density matrix of the microstates, we average over them.

S_{fine} **Fine-grained (von-Neumann) entropy:** It is the von-Neumann entropy of quantum mechanics, sometimes also called entanglement entropy. Consider a state for which we know the density matrix ρ of the microstates, and compute

$$S_{\text{fine}} = \mathrm{Tr}\left(\rho \log \rho\right) \ . \tag{6.12}$$

This is the fine-grained von Neumann entropy of ρ. If the von Neumann entropy is zero, then the microstates are in a pure state. Therefore it measures the mixedness of the microstates.

The entropy of a black hole, as computed by (6.8), is of the type S_{coarse}. Indeed, we computed it without knowing the precise black hole microstates, and it is a thermodynamic entropy since it obeys the first law of thermodynamics. On the other hand, the entropy of the black hole computed by (6.10) is of the type S_{fine} since it is computed from the partition function. In turn, this implies that the specific computation of the black hole entropy through (6.10) has information on the black hole microstates.
Since the entropy (6.10) is independent of the cut-off area A_{box}, it is proportional to the horizon area A_{hor}, and should have information on the black hole microstates, Wallace (2019) concludes that the microscopic degrees of freedom of the black hole should be localised at the horizon. In this way, the phenomenological membrane **(SH)** can be seen as the emergent description of the fundamental quantum membrane **(QMP)** at the horizon, where the degrees of freedom of the black hole, responsible for its thermodynamics, are localised. In the following sections, we are going to analyse such a conclusion in light of recent results on black hole physics.
The next section is a conceptually oriented presentation of new results on black hole physics obtained by studying the gravitational path integral in Penington et al. (2020) and Almheiri et al. (2020b). The upshot of their work is the derivation, directly from the gravitational path integral, of a formula, which has been called in the literature *the island rule*, for computing the entropy of the Hawking radiation. According to this new rule, Hawking radiation carries some information on the black hole interior so that the black hole evaporation process is unitary and firewall-less.

6.3 A Modern Perspective on Black Holes

This section is a conceptually oriented presentation of new results on black hole physics obtained by studying the gravitational path integral in Penington et al. (2020) and Almheiri et al. (2020b). The upshot of their work is the derivation, directly from the gravitational path integral, of a formula, which has been called in the literature *the island rule*, for computing the entropy of the Hawking radiation. According to this new rule, the Hawking radiation carries some information on the black hole interior so that the black hole evaporation process is unitary and firewall-less.
In this section, I first sketch the derivation of the formula for computing the entropy of an evaporating black hole derived from the gravitational path integral, showing the main difference with Hawking (1975)'s result. I then present the exterior description of the black hole emerging in this new approach: what has been called the *central dogma*. Finally, I present the island rule for computing the entropy of the Hawking radiation, and *entanglement wedge reconstruction*, i.e. the description of the interior of a black hole.

6.3.1 The Gravitational Path Integral

The starting point of Penington et al. (2020) and Almheiri et al. (2020b) is the same as the one developed in section §6.2: the path integral (6.9), which describes the partition function of a Schwarzshild black hole. The problem concerns the computation of the partition function, and of the entropy of the black hole from the partition function, using the saddle-point approximation.[11] A useful method for computing the von-Neumann entropy $S_{\text{BH}} = -\text{Tr}\,[\rho \log \rho]$ of the black hole from the partition function is called the *replica trick*. It consists in first computing the Rényi entropy $S_n = \text{Tr}\,[\rho^n]$ for integer n, then performing an analytic continuation for real n, and finally computing the von-Neumann entropy as

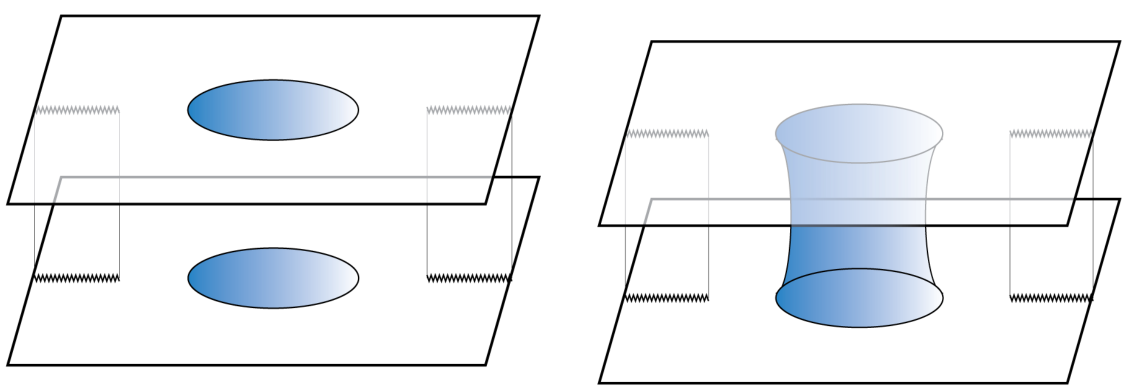
$$S_{\text{VN}} = \lim_{n\to 1} \frac{1}{n-1} S_n \ . \tag{6.13}$$

This method has been used for example by Calabrese and Cardy (2009) for computing the entanglement entropy in two dimensional conformal field theory, with outstanding

[11] Note that in order for the saddle-point approximation to be reliable, one needs specific conditions on the matter content. This is discussed in Almheiri et al. (2020b).

results. In this setup, the computation of $\mathrm{Tr}[\rho^n]$ can be seen as the computation of a single observable $\mathrm{Tr}\rho$ in n copies (or replicas) of the original system, choosing appropriate boundary conditions that connect the various replicas.

Without delving into unnecessary details, the upshot of the work of Penington et al. (2020) and Almheiri et al. (2020b) is the following: there are two different saddle-points that contribute to the computation of $\mathrm{Tr}[\rho^n]$ and therefore of S_{VN}. One of them is the saddle-point considered by Hawking (1975), in which the different replicas are sewn together along with their branch points, as is usually done in Quantum Field Theory calculations (see figure 6.3a). There is, however, another saddle-point where gravity dynamically glues together the different replicas through a wormhole geometry (see figure 6.3b). These new geometrical structures responsible for the existence of a new saddle-point of the path integral (6.9) are called *replica wormholes*.



(a) Hawking saddle

(b) Replica wormhole saddle

Figure 6.3: Figure 6.3a represents the Hawking saddle, in which the different replicas (white planes) are sewn together along the branch points (wiggly lines). Figure 6.3b represents the replica wormhole saddle, in which there is a replica wormhole (blue cylinder) dynamically gluing together different replicas (white planes).

It is important to understand that when we speak of replica wormholes, we are not referring to physical wormholes that exist in spacetime like the wormholes that we may typically imagine. These mathematical objects do not live in spacetime and do not have a physical meaning, as it was for the wormholes of the eternal black hole described in section §5.1.3. Instead, replica wormholes are a mathematical construct connecting two different replicas. The question of how this abstract geometric construction relates to real physical objects is open, and, in my opinion, it should be pursued. Understanding the true nature of these mathematical objects and how they might relate to real physical phenomena is a fascinating and challenging question central to our understanding of quantum gravity and the nature of space and time.

These new connections between different replicas, absent in the semiclassical Hawking calculation, have higher-order topology and are non-perturbatively small. Before the Page time, the new saddles are heavily suppressed, and the Hawking saddle dominates the path integral. After the Page time, however, the new replica wormhole saddle becomes important and contributes significantly to the final result of the entropy. The consequence of the presence of the new saddle is that the von-Neumann entropy of the black hole can be computed as

$$S_{\text{BH}} = \min_{\chi} \left\{ \text{ext}_{\chi} \left[\frac{A(\chi)}{4G_N \hbar} + S_{\text{semi-cl}}(\Sigma_{\chi}) \right] \right\} , \tag{6.14}$$

where χ is a quantum extremal surface, Σ_{χ} is the region between χ and the cut-off surface of the gravitational path integral, $A(\chi)$ is the area of the quantum extremal surface χ and $S_{\text{semi-cl}}(\Sigma_{\chi})$ is the von Neumann entropy of quantum fields on Σ_{χ}. The quantity in square brackets is called *generalised entropy* $S_{\text{gen}}(\chi)$ of the black hole. The computation of S_{BH} consists in first finding all the surfaces χ which extremise the generalised entropy, and then choosing the quantum extremal surface χ which gives the minimal generalised entropy.

The formula (6.14) was first derived in the context of the AdS/CFT correspondence (Ryu and Takayanagi (2006); Hubeny et al. (2007); Faulkner et al. (2013)).[12] The same formula (6.14) has been used, under the name of *QES prescription*, by Penington (2019) in the context of an evaporating black hole showing that, according to (6.14), the black hole evaporation process is unitary and firewall-less.[13] The main advance of Penington et al. (2020) and Almheiri et al. (2020b) is that they derived (6.14) directly from the gravitational path integral, without resorting to the AdS/CFT correspondence. Therefore, these works suggest a way out of the black hole information paradox since they derived, from the path integral approach to gravity, the conjectured formula (6.14) which guarantees unitary and firewall-less black hole evaporation process.

The main conclusion of this section is that black holes have an entropy, which is given by (6.14), and they can therefore be considered thermodynamic objects in the fullest

[12]The AdS/CFT correspondence was originally introduced in Maldacena (1999a). For a philosophical discussion on the AdS/CFT correspondence see De Haro et al. (2016).

[13]For a conceptually oriented presentation of Penington (2019) and for the consequences of the resolution of the black hole information paradox on semiclassical physics see Cinti and Sanchioni (2021a).

sense, with a statistical mechanical underpinning. Indeed, we derived the black hole entropy (6.14) from the black hole partition function (6.9). However, the role played by **(QMP)** is not clear anymore since, in general, the quantum extremal surface χ is not localised at the event horizon of the black hole, and it is not clear how the quantum extremal surface χ is related to **(QMP)**. We are going to analyse the physical implication of (6.14) and its bearing on **(QMP)** in the following sections.

6.3.2 Central Dogma, a.k.a. the Exterior Description

The fact that black holes thermodynamics has a statistical mechanical underpinning is a strong motivation for interpreting them as ordinary quantum systems, with N degrees of freedom given by $N \approx \log S_{BH}$, obeying ordinary quantum mechanical rules. In particular, N would represent the dimension of the Hilbert space, which describes the black hole. The proposal of interpreting black holes as ordinary quantum systems is the essence of the *central dogma*:

(CD) Central Dogma: As seen from the outside, a black hole can be described in terms of a quantum system with $\frac{A}{4G_N}$ degrees of freedom, which evolves unitarily under time evolution.

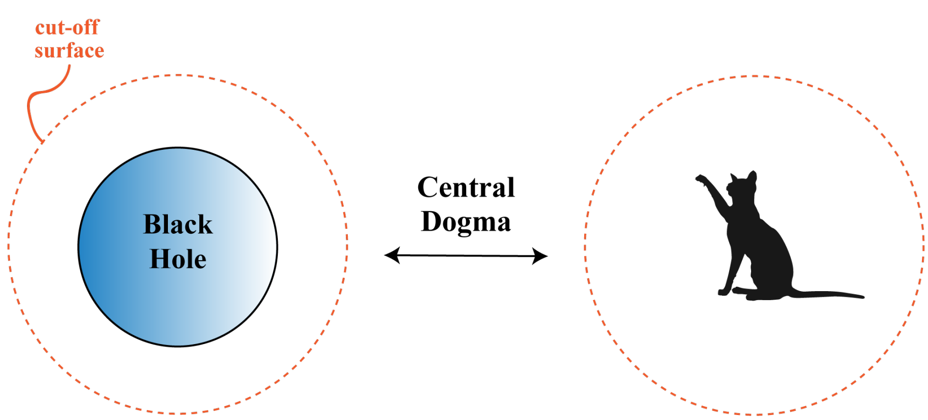


Figure 6.4: The Central Dogma. There is a one-to-one correspondence between a system of a black hole with its surroundings (up to the cut-off surface, represented by the orange dashed line) and a quantum mechanical system (pictorially represented by a Schrödinger cat).

The name Central Dogma comes from biology, where it refers to the information transfer from DNA and RNA to proteins. Here the statement of the central dogma is also about (quantum) information, the information transfer from the quantum system to the gravitational system. It is important to note that, in the definition of **(CD)**, the black hole degrees of freedom are given by $\frac{A}{4G_N\hbar}$ and not by $\frac{A_{\text{hor}}}{4G_N\hbar}$. Indeed, **(CD)** states that the black hole and *the whole spacetime around it* up to a certain surface of area A, can be replaced by a quantum system, with $\frac{A}{4G_N\hbar}$ degrees of freedom, obeying the standard rules of quantum mechanics. Therefore, the main message of **(CD)** is that the whole system of the black hole and its surroundings are together an ordinary quantum system.

Note that the degrees of freedom of the central dogma are not manifest in the gravity description. **(CD)** states that the gravitational system composed by the black hole and the matter located in its exterior, up to a cut-off surface of area A, is *equivalent* to a quantum mechanical system with N degrees of freedom. From **(CD)** it is not clear what those degrees of freedom are, and neither where they are located. In the AdS/CFT language, we would say that the gravitational system is *dual* to a quantum mechanical system living at the boundary of the AdS spacetime. In this more general case, independently from AdS/CFT, we say that these two systems are equivalent: the gravitational system of the black hole and its surroundings is equivalent to the quantum mechanical system with $\frac{A}{4G_N\hbar}$ degrees of freedom. In other words, we do not have any insights from **(CD)** on the location of those degrees of freedom in the gravity description. In particular, we do not have any clues that those degrees of freedom are located at the horizon. Indeed, since the number of degrees of freedom is proportional to A and not to A_{hor}, at best we have clues that they are not located at the horizon.

In conclusion, **(CD)** gives a picture for the black hole as seen from the outside within the gravitational path integral description. Up to this point, nothing has been said about the interior structure of the black hole. In particular, we do not know if the degrees of freedom of **(CD)** also describe the interior of a black hole. In the next section, we are going to review the interior description of the black hole emerging from the path integral, and in particular from the implementation of the formula (6.14).

6.3.3 Entanglement Wedge Reconstruction, a.k.a. the Interior Description

Understanding how an observer located outside the black hole can describe the interior would be of extreme importance for our analysis since, if we want to analyse the status of **(QMP)**, we need to understand which degrees of freedom describe the black hole interior. Since **(QMP)** and **(CD)** represent the same degrees of freedom, i.e. the microstates of the black hole as seen by an exterior observer, one would like to find evidence for **(QMP)** within the description given by **(CD)**. Thus, we need to understand if the degrees of freedom of **(CD)** describe the interior of the black hole or not. In order to do it, let me review the entanglement wedge structure of an evaporating black hole, presented in §3.4.2.

Let us start by considering how to implement the formula (6.14) in the case of an evaporating black hole. The first task for computing the black hole entropy is to find a quantum extremal surface χ which minimises the generalised entropy $S_{\text{gen}}(\chi)$. This minimal quantum extremal surface has been called the *quantum Ryu-Takayanagi* surface (Hubeny et al. (2007)).[14] For an evaporating black hole there are two quantum extremal surfaces: the empty surface $\varnothing$ and a non-vanishing surface $\tilde{\chi}$ which lies just inside the event horizon. At early times in the black hole evaporation process, i.e. before a time t_P called the Page time,[15] the generalised entropy of the surface $\varnothing$ is smaller than the generalised entropy of the surface $\tilde{\chi}$. Therefore, $\varnothing$ is the quantum Ryu-Takayanagi surface, and the black hole entropy is $S_{\text{semi-cl}}\left(\Sigma_{\varnothing}\right)$, which is Hawking's original result. However, at late times in the black hole evaporation process, the generalised entropy of the surface $\tilde{\chi}$ becomes larger than the generalised entropy of the surface $\varnothing$, which means that $\tilde{\chi}$ is the new quantum Ryu-Takayanagi surface. In other words, at the Page time we have a phase transition between the two surfaces $\varnothing$ and $\tilde{\chi}$. At late times the entropy of the black hole is approximately given by $\frac{A_{\text{hor}}(\tilde{\chi})}{4G_N\hbar}$, which is equal to the thermodynamic entropy of the black hole. In conclusion, the entropy of the black hole is $S_{\text{semi-cl}}\left(\Sigma_{\varnothing}\right)$ before the Page time and $\frac{A_{\text{hor}}(\tilde{\chi})}{4G_N\hbar}$ after the Page time, i.e. it is given by

$$S_{\text{BH}} = \min\left\{ S_{\text{semi-cl}}\left(\Sigma_{\varnothing}\right), \frac{A_{\text{hor}}\left(\tilde{\chi}\right)}{4G_N\hbar} \right\}, \tag{6.15}$$

which entails that the black hole evaporation process follows the Page curve of figure 6.5.

The entropy (6.15) we have just computed for the evaporating black hole is of the type S_{fine}, and it should thus have some information on the degrees of freedom of **(CD)**.

[14]For philosophical discussion on the quantum Ryu-Takayanagi see Bain (2020b).

[15]The Page time is the time at which the semiclassical entropy of the Hawking radiation becomes larger than the black hole's thermodynamic entropy. From now on, I will use the terminology early (late) time and before (after) the Page time interchangeably.

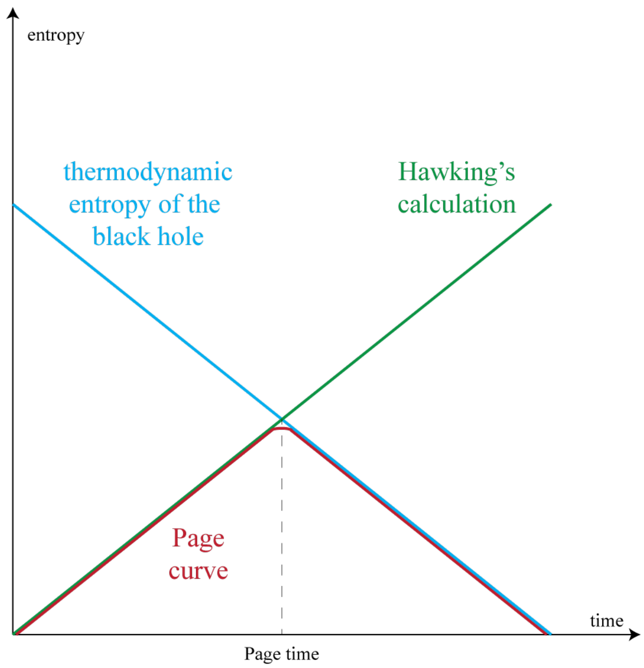


Figure 6.5: The Page curve of an evaporating black hole. The light blue line represents the thermodynamic entropy of the black hole; the green line represents the entanglement entropy of the black hole in the original Hawking calculation; the red line represents the Page curve, which is the fine-grained entropy of the black hole computed by (6.14).

The entropy (6.15) has been calculated combining the geometrical information (area term of (6.14)) and the state of quantum fields (entanglement term of (6.14)) included in the region bounded by the quantum Ryu-Takayanagi surface χ (which is $\varnothing$ at early times and $\tilde{\chi}$ at late times). For this reason, it seems natural to think that the degrees of freedom of **(CD)** describe the physics which takes place within the region bounded by the minimal surface χ, i.e. the description of black hole system and its surroundings of an exterior observer. In order to make this intuition more precise, it is useful to introduce the notion of the *entanglement wedge* of the quantum Ryu-Takayanagi surface χ, which is the region of spacetime bounded by the cut-off surface and the quantum Ryu-Takayanagi surface χ. The fact that the degrees of freedom of **(CD)** describe the physics up to the quantum Ryu-Takayanagi surface, an idea supported by the entropy computation of the black hole from the path integral, is the second main hypotheses of this section. It is called the entanglement wedge reconstruction conjecture:[16]

[16]There are many consistency checks of this hypothesis, especially within the AdS/CFT correspondence (Dong et al. (2016); Jafferis et al. (2016)).

(EWR) **Entanglement wedge reconstruction** says that all physical quantities in the entanglement wedge of a Ryu-Takayanagi surface χ are represented by operators in the quantum system which represent that spacetime region (think for example of how **(CD)** encodes information regarding the spacetime region accessible to an exterior observer).

(EWR) has some interesting consequences on the structure of black holes. Before the Page time, the Ryu-Takayanagi surface is the empty surface $\varnothing$ and the entanglement wedge is just the region inside the cut-off surface (blue region of figure 6.6a). However, at late times in the black hole evaporation process, the Ryu-Takayanagi surface $\tilde{\chi}$ lies inside the black hole, which means, by **(EWR)**, that the degrees of freedom of **(CD)** now describe just a portion of the interior of the black hole because just a portion of the black hole interior lies within the entanglement wedge of $\tilde{\chi}$ (blue region of figure 6.6b). A question we could ask, which is relevant for understanding the status of **(QMP)**, is which system describes the other part of the black hole interior at late times.
In order to address this issue, let us consider the entropy of the Hawking radiation. In this setup, the Hawking radiation is described as the complement system of the central dogma. The formula for computing the entropy of Hawking radiation, that can be derived from the gravitational path integral, is similar to (6.14). The main difference is that the region Σ_χ can be disconnected. In particular, Σ_χ can be the union of two disconnected regions: Σ_{rad}, which is the region outside the cut-off surface, and Σ_{Island} which is the region between the origin of the coordinate system and the quantum extremal surface χ. The formula for computing the entropy of the Hawking radiation is given by:

$$S_{\text{rad}} = \min_\chi \left\{ \text{ext}_\chi \left[\frac{\text{Area}\,(\chi)}{4G_N\hbar} + S_{\text{semi-cl}}\left(\Sigma_{\text{rad}} \cup \Sigma_{\text{Island}}\right) \right] \right\} \ . \tag{6.16}$$

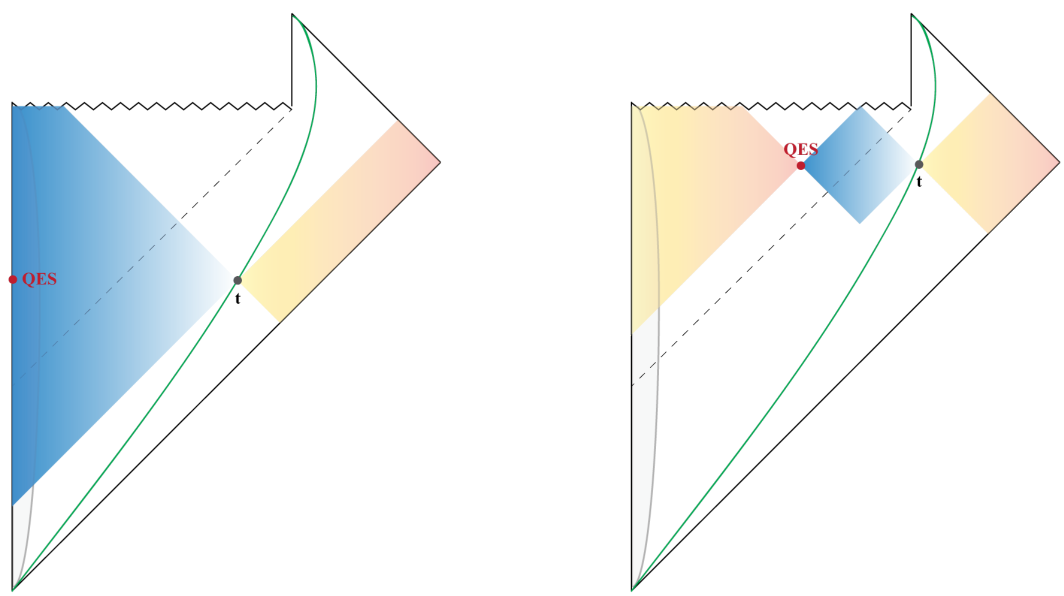
Before the Page time, the quantum Ryu-Takayanagi is $\varnothing$, which corresponds of the situation in which there is no island, i.e. $\Sigma_{\text{Island}} = \varnothing$. In this case the fine-grained entropy of the Hawking radiation computed by (6.16) gives the same result of Hawking's original calculation, i.e. $S_{\text{rad}} = S_{\text{semi-cl}}\left(\Sigma_{\text{rad}}\right)$. However, after the Page time, the quantum extremal surface is $\tilde{\chi}$, there is a non-zero island contribution, and the entropy of the radiation is given by $\frac{A_{\text{hor}}(\tilde{\chi})}{4G_N\hbar}$.[17] The fine-grained entropy of the Hawking radiation, as

[17]The entanglement entropy contribution in S_{gen} is tiny after the Page time since the island contains much of the interior Hawking modes which purify the outgoing radiation.

computed by (6.16) is

$$S_{\text{rad}} = \min\left\{ S_{\text{semi-cl}}\left(\Sigma_{\varnothing}\right), \frac{A_{\text{hor}}\left(\tilde{\chi}\right)}{4G_N\hbar} \right\} \quad , \tag{6.17}$$

and follows the Page curve of figure 6.5. Before the Page time, the entanglement wedge of the radiation is a region in the exterior of the cut-off surface (see the orange region of figure 6.6a), while after the Page time, it is the union of a region outside the cut off-surface and the island (see the orange region of figure 6.6b).



(a) Before the Page time

(b) After the Page time

Figure 6.6: The Penrose diagrams of an evaporating black hole. The red dot represents the location of the Ryu-Takayanagi surface, the green line represents the cut-off surface, the dashed line represents the event horizon, the blue region represents the entanglement wedge of the degrees of freedom of **(CD)**, and the orange region represents the entanglement wedge of the Hawking radiation.

Figure (6.6) gives us a clear picture of the quantum systems which describe the interior of an evaporating black hole. Before the Page time, the degrees of freedom of **(CD)** describe all the black hole interior and the surroundings of the black hole up to a cut-off surface. After the Page time, however, the degrees of freedom of **(CD)** describe only a portion of the interior. Indeed, the other portion of the interior is described by Hawking radiation since part of the black hole interior belongs to the entanglement wedge of the radiation.

Before concluding, let me underline another important feature of the black hole spacetime structure emerging from the works of Penington et al. (2020) and Almheiri et al. (2020b).[18] In ordinary General Relativistic black holes, the interior and the exterior of the black hole are two disconnected regions of spacetime. However, this is not the case when considering quantum effects, as developed in this section. Indeed, after the Page time, it is, in principle, possible to make complex operations on the Hawking radiation (accessible to an asymptotic observer) to recover information on the interior of the black hole since a portion of the interior and the Hawking radiation belong to the same quantum system. To be more precise, this is precisely an instance of the violation of **(SD)** discussed in chapter §5. In this framework, even if the interior and the exterior of a black hole are spacelike related for a fixed time slice, there are non-trivial connections between the Hawking radiation (which is in the exterior of the black hole) and the interior (the island), where these non-trivial connections come from the fact that the radiation and the interior do not commute since they belong to the same quantum system. This are, indeed, particular manifestations of we have called in chapter §5 the *generalised causal structure*. After the Page time, relations such as those relevant for the generalised causal structure become important to ensure unitarity, and there is, thus, a clear violation of **(SD)**.

6.4 Do We (really) Need (QMP)?

I provide now four arguments which show that **(QMP)** is not a valid description of black hole physics at the semiclassical level, i.e. that no membrane of black hole microstates should be posited at the horizon. From the point of view of the results presented in §6.3, **(QMP)** seems to be:

(i) **Unnecessary. (QMP)** was first introduced to have a statistical mechanical underpinning of black hole thermodynamics. However, from the point of view of section §6.3, one can understand the statistical mechanics of black holes as a simple consequence of **(CD)**. Indeed, **(CD)** states that the black hole system, together with its surroundings up to the cut-off surface, is an ordinary quantum mechanical system with $N = \frac{A}{4G_N\hbar}$ degrees of freedom. If this is the case, then, the black hole system carries an entropy given by $S = \log N$ as any ordinary quantum

[18] For an in-depth discussion of this topic, which I only sketch here, see Cinti and Sanchioni (2021a).

mechanical system. Therefore accepting **(CD)** entails, without any further assumptions, a statistical mechanical underpinning of black holes thermodynamics, because the black hole is a statistical mechanical system thanks to **(CD)**. For this reason, positing a quantum membrane at the horizon is unnecessary.

(ii) **Conventional.** Even though **(QMP)** is not necessary to have a statistical mechanical underpinning of black holes thermodynamics, it could be the case that **(QMP)** is a logical consequence of **(CD)** itself. In other words, it could be that the degrees of freedom of **(CD)** are located at the event horizon, thus forming the sought-after quantum membrane. However, the degrees of freedom of **(CD)** are not explicitly manifest in the gravitational description, and the path-integral does not tell us anything about their location. **(CD)** just tells us that the black hole system is equivalent to a quantum mechanical system, nothing more. Moreover, the hypothesis that the black hole degrees of freedom are located at the horizon was advanced because, in the old Hawking calculation, the entropy of the black hole was proportional to the horizon area as described by equation (6.10). However, according to the more careful derivation of black hole entropy provided in §6.3, the black hole's entropy is not proportional to the horizon area but is given by the generalised entropy $S_{\text{gen}}(\chi)$, where χ is the quantum Ryu-Takayanagi surface which in general does not correspond to the event horizon, as evident from equation (6.14). Therefore, there is no physical reason to locate the black hole degrees of freedom at the black hole horizon. This choice seems to be purely conventional since we do not have any clues that this is the case from the derivation presented in §6.3.
However, nothing prevents us from guessing the location of those degrees of freedom. In particular, one might think that the degrees of freedom of **(CD)** are coming from the thermal atmosphere in the neighbourhood of the black hole, a "brick wall" at some Planck distance from the horizon. This "brick wall" could be understood as the quantum membrane of **(QMP)**. However, nothing in the physical description justifies this guessing, which thus seems fictitious.

(iii) **Incomplete.** Even though **(QMP)** is unnecessary and purely conventional, one could still think that the degrees of freedom of **(CD)** are located at the event horizon. However, this seems awkward from the point of view of the interior reconstruction presented in §6.3. Indeed, before the Page time, the degrees of freedom of **(CD)** are not describing just the black hole, but also the whole spacetime around it up to the cut-off surface (see the blue region in figure 6.6a). On the other hand, after the Page time, the degrees of freedom of **(CD)** are describing just a portion of the black hole interior since the other portion belongs to the entanglement wedge of the Hawking radiation (see the blue region in figure 6.6b). Thus, if the degrees of freedom of **(CD)** were located at the horizon, and were meant to describe the physics of the black hole, which are the two core features of **(QMP)**, after the Page time the description they provide would be incomplete.

(iv) **Inconsistent.** Moreover, we have seen at the end of the previous section that a peculiar feature of the resolution presented in §6.3 is that there are non-local connections between the Hawking radiation (accessible to an exterior observer) and the black hole interior. In particular, it is possible, in principle, to perform complex operations on the Hawking radiation to extract information on the black hole interior. If a membrane is posited at the black hole horizon, as required by **(QMP)**, these new connections, which are responsible for the unitarity of the black hole evaporation process, would be spoiled. Indeed, take for example the "brick wall" model presented above. There, the presence of the membrane at the horizon would not permit correlations between the island and the Hawking radiation. Since the "brick wall" is thermal, the interior and the exterior of the black hole are disconnected, and there cannot be any connections, not even entanglement, across the black hole horizon. In particular, there cannot be the non-trivial connections which are crucial to have the island and the radiation belong to the same quantum system. Therefore, **(QMP)** is inconsistent with an important feature of the black hole picture presented in §6.3.

In conclusion, a membrane posited at the horizon can only have at-best phenomenological import. From the point of view of the new results on black hole physics obtained in Penington et al. (2020) and Almheiri et al. (2020b), the quantum membrane paradigm should be abandoned. In other words, if a membrane is introduced for describing a black hole as seen from the outside, it should only be taken as a phenomenological crutch, not as a physical structure.

6.5 Conclusions and Outlooks

In this section, I analysed the status of **(QMP)** in light of recent results on black hole physics, which I reviewed in section §6.3. I argued that, if one is willing to preserve the equivalence principle of General Relativity also at the semiclassical level, the quantum membrane cannot be a fundamental feature of black holes. However, it could still be the case that the quantum membrane is a valid candidate for the description of a black hole in a particular reference system, i.e. the exterior description. Concerning this last possibility, I put forward four arguments ruling out the possibility of having a membrane of black hole microstates at the horizon. Assuming both **(CD)** and **(EWR)** to be reasonable hypotheses for the description of the physics of black holes, positing the quantum membrane at the horizon is:

- *conventional*, since nothing in the path integral description of the black hole, tells us where the degrees of freedom of **(CD)** are located;
- *unnecessary*, since giving a statistical mechanical underpinning of black hole thermodynamics is a consequence of **(CD)**;
- *incomplete*, since, if one assigns the degrees of freedom of **(CD)** to a quantum membrane at the horizon, after the Page time that membrane will describe only a portion of the interior;
- *inconsistent*, since positing a quantum membrane at the horizon would spoil the non-local connections between the interior and the exterior crucial of **(EWR)**.

However, despite arguing against the membrane paradigm, I did not offer any alternative to it. It will be an essential task for future works to investigate the status of the fundamental ontology underlying the work of Penington et al. (2020) and Almheiri et al. (2020b).

Moreover, the features presented in section §6.3 were not features of the black hole itself, but the consequences of the presence of a horizon in spacetime. It is well known that also de Sitter spacetime has a horizon, the cosmological horizon. Therefore, this work could have some critical consequences on our understanding of the big bang and of cosmology.

Furthermore, the results of Penington et al. (2020) and Almheiri et al. (2020b) seem

to rely on two main conceptual moves. One of them is the violation of Spacetime Distinctness, i.e. the fact that the interior and the exterior of the black hole are non-locally connected. The other one is the fact that non-perturbative corrections (the replica wormholes of section §5.1) should be added to the semiclassical description in order to have unitary black hole evaporation. It would be of great interest to better characterise the overall conceptual strategy behind this new perspective on black hole physics.[19]
Finally, the results of this section should be quite general since they are based on features of the gravitational path integral. For this reason, it would be interesting to compare the status of **(QMP)** in other theories of quantum gravity, such as that Loop Quantum Gravity or Asymptotic Safety.

[19]This issue will be discussed in detail in a forthcoming paper.

Chapter 7

Conclusions

In this thesis, I have provided a philosophical discussion of black hole paradoxes. Through an in-depth examination of these paradoxes, I have aimed to shed light on the underlying issues and inconsistencies in our understanding of black holes. The main focus of this thesis has been to identify where the inconsistencies lie in the various paradoxes and to point out the implicit assumptions that are made in the physics reconstruction of these paradoxes. One of this thesis's main results is identifying the causal structures that underlie the various paradoxes. By analysing the causal structures, it has been possible to pinpoint where the inconsistencies lie in the paradoxes and to understand the underlying issues more clearly. Another key result of this thesis is the pointing out of implicit assumptions in the physics reconstruction of the paradoxes, such as spacetime distinctness and semiclassical exactness. These assumptions are problematic and contribute to the inconsistencies in our understanding of black holes. In addition, this thesis has provided a taxonomy of possible black hole paradox resolution and matched it with the relevant literature in physics. This has helped to contextualise the various proposed solutions and to understand how they relate to one another. Furthermore, this thesis has discussed what the violation of spacetime distinctness and semiclassical exactness implies on our image of black holes. It has been shown that these assumptions play a crucial role in shaping our understanding of black holes and that their violation has far-reaching implications. One of the most significant results of this thesis is the demonstration that the resolution of the AMPS and AMPSS paradoxes recently proposed relies on the same conceptual move: introducing the generalised causal structure (which is not just a naive superposition of entanglement and relativistic spacetime). This is a crucial step in resolving these paradoxes and understanding the underlying issues more clearly. Finally, this thesis has begun a discussion on what the more conservative resolution of the AMPS and AMPSS paradox implies on the ontology of black holes. However, it is essential to note that this is just the beginning of an inquiry into this vast subject, and further research is needed to fully understand these resolutions' implications.

In this thesis, I have provided a philosophical discussion of black hole paradoxes and aimed to shed light on the underlying issues and inconsistencies in our understanding of black holes. However, despite the insights and contributions made in this thesis, there are still open problems that need to be addressed in future research.

One of the main open problems is the need to defend route (iii) as the preferred resolution of black hole paradoxes against the competitor resolutions. In the thesis, I have broadly spoken about the fact that it is the more conservative resolution since it does not give up well-known physical principles, such as the equivalence principle and unitarity, in a regime where we think they should apply. However, I still need to put forward a full defence of it. This is an important project that needs to be undertaken in the future. A full defence of the route (iii) resolution would require a detailed examination of its implications and a comparison to the other resolutions proposed in the literature. Additionally, it would be necessary to examine the experimental and observational evidence that could support or contradict the route (iii) resolution.

Another open problem is the need to further investigate the ontology of black holes in quantum gravity. In the thesis, I have just begun an investigation on this subject, and it is essential to develop a positive ontology for black holes in light of recent results. A starting point would be to provide a taxonomy of all possible black hole definitions. Classically, this is already a complex issue, as black holes can be thought of as "what is inside the event horizon," but they are also global solutions, which creates problems. These problems are even more severe when considering the picture of black holes in quantum gravity from the "route (iii) resolution of black hole paradoxes." Before the Page time, the degrees of freedom associated with the black hole is inside the event horizon and a small part of its surrounding. After the Page time, however, the degrees of freedom related to the black hole is only a part of the interior (the other part, the island, is associated with Hawking radiation). This creates problems in defining a black hole in quantum gravity, as it looks pretty different from the classical picture, and it seems to change over time. Developing a consistent view of black holes in these regards would require a deep understanding of the implications of the route (iii) resolution on the ontology of black holes. It would also be essential to examine the implications of this new ontology on other areas of physics, such as information theory, thermodynamics and quantum mechanics.

In conclusion, while this thesis has made significant contributions to understanding black hole paradoxes, I would need to defend route (iii) as the preferred resolution of black hole paradoxes and the need to further investigate the ontology of black holes in quantum gravity. Resolving these open problems will be crucial to fully understanding the nature of black holes and their role in our understanding of the universe.

Beyond the immediate outlooks of this thesis, I would like to explore the relationship between black hole paradoxes and the measurement problem in quantum mechanics in the near future. The ongoing debate between determinism and indeterminism in the history and philosophy of physics is a longstanding and important discussion. I aim to provide a novel perspective on this classic issue by showing that determinism and indeterminism are not polar opposites but different perspectives on a richer and more complex physical reality.

To achieve this, I plan to follow a single unified thread, black hole complementarity and analyse the conceptual connections between black holes in quantum gravity and the measurement problem in quantum mechanics. As I have demonstrated throughout this thesis, black hole paradoxes, in one form or another, concern the incompatibility between a fine-grained deterministic description of black holes in terms of statistical mechanics associated with their microscopic structure and the coarse-grained indeterministic thermodynamic description of the radiation emitted by the black hole as it evaporates. This takes the form of a conflict between determinism and indeterminism within the physics of quantum black holes.

Similarly, the measurement problem in quantum mechanics is a consequence of the incompatibility between the deterministic dynamics governing the ordinary evolution of quantum states, encoded in the Schrödinger equation, and the indeterministic probabilistic laws governing measurement interactions, most notably in the form of wave function collapse and the Born rule. The measurement problem in quantum mechanics is still unsolved. Still, the analogous situation in the context of black holes has a well-accepted solution, at least among string theorists, dubbed in this thesis "route (iii)", which is based on ER=EPR-like phenomena and connected to the black hole complementarity program. Black hole complementarity, at its core, is the claim that the deterministic interior, including the statistical mechanical description of the black hole, is equivalent, yet complementary, to the thermodynamic description of the exterior encoded in the radiation's indeterministic behaviour. I aim to deeply understand the analogies and differences between these two complementarity proposals. Moreover, I would like to understand why complementarity is widely shunned as a solution to the measurement problem in quantum mechanics and whether or not insights coming from black hole physics and the black hole version of complementarity can be used to solve foundational issues in quantum mechanics. This would require a detailed examination of the similarities and differences between the measurement problem in quantum mechanics and the problem of black hole paradoxes, as well as an analysis of the implications of black hole complementarity for the foundations of quantum mechanics.

Bibliography

Abbott, B. P., R. Abbott, T. Abbott, M. Abernathy, F. Acernese, K. Ackley, C. Adams, T. Adams, P. Addesso, R. Adhikari, et al. (2016). Observation of gravitational waves from a binary black hole merger. *Physical review letters 116*(6), 061102.

Akers, C., N. Engelhardt, D. Harlow, G. Penington, and S. Vardhan (2022). The black hole interior from non-isometric codes and complexity. *arXiv preprint arXiv:2207.06536*.

Akers, C. and G. Penington (2022). Quantum minimal surfaces from quantum error correction. *SciPost Physics 12*(5), 157.

Almheiri, A. (2018). Holographic quantum error correction and the projected black hole interior. *arXiv preprint arXiv:1810.02055*.

Almheiri, A., X. Dong, and D. Harlow (2015, Apr). Bulk locality and quantum error correction in ads/cft. *Journal of High Energy Physics 2015*(4).

Almheiri, A., N. Engelhardt, D. Marolf, and H. Maxfield (2019a, Dec). The entropy of bulk quantum fields and the entanglement wedge of an evaporating black hole. *Journal of High Energy Physics 2019*(12).

Almheiri, A., N. Engelhardt, D. Marolf, and H. Maxfield (2019b). The entropy of bulk quantum fields and the entanglement wedge of an evaporating black hole. *Journal of High Energy Physics 2019*(12), 1–47.

Almheiri, A., N. Engelhardt, D. Marolf, and H. Maxfield (2019c). The entropy of bulk quantum fields and the entanglement wedge of an evaporating black hole. *JHEP 12*, 063.

Almheiri, A., T. Hartman, J. Maldacena, E. Shaghoulian, and A. Tajdini (2020a). The entropy of hawking radiation.

Almheiri, A., T. Hartman, J. Maldacena, E. Shaghoulian, and A. Tajdini (2020b). Replica wormholes and the entropy of hawking radiation. *Journal of High Energy Physics 2020*(5), 13.

Almheiri, A., T. Hartman, J. Maldacena, E. Shaghoulian, and A. Tajdini (2020c). Replica Wormholes and the Entropy of Hawking Radiation. *JHEP 05*, 013.

Almheiri, A., T. Hartman, J. Maldacena, E. Shaghoulian, and A. Tajdini (2021). The entropy of hawking radiation. *Reviews of Modern Physics 93*(3), 035002.

Almheiri, A., R. Mahajan, J. Maldacena, and Y. Zhao (2020a, Mar). The page curve of hawking radiation from semiclassical geometry. *Journal of High Energy Physics 2020*(3).

Almheiri, A., R. Mahajan, J. Maldacena, and Y. Zhao (2020b, Mar). The page curve of hawking radiation from semiclassical geometry. *Journal of High Energy Physics 2020*(3).

Almheiri, A., R. Mahajan, J. Maldacena, and Y. Zhao (2020c). The Page curve of Hawking radiation from semiclassical geometry. *JHEP 03*, 149.

Almheiri, A., D. Marolf, J. Polchinski, D. Stanford, and J. Sully (2013). An Apologia for Firewalls. *JHEP 09*, 018.

Almheiri, A., D. Marolf, J. Polchinski, and J. Sully (2013). Black Holes: Complementarity or Firewalls? *JHEP 02*, 062.

Ammon, M. and J. Erdmenger (2015). *Gauge/Gravity Duality: Foundations and Applications* (1st ed.). USA: Cambridge University Press.

Armas, J., N. A. Obers, and M. Sanchioni (2016). Gravitational tension, spacetime pressure and black hole volume. *Journal of High Energy Physics 2016*(9), 124.

Bain, J. (2020a). The rt formula and its discontents: spacetime and entanglement. *Synthese*, 1–28.

Bain, J. (2020b). Spacetime as a quantum error-correcting code? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 71*, 26–36.

Balasubramanian, V., A. Kar, and T. Ugajin (2021). Islands in de sitter space. *Journal of High Energy Physics 2021*(2), 1–32.

Baron, S. and B. L. Bihan (2022). Spacetime quietism in quantum gravity. Forthcoming in Antonio Vassallo (ed.), The Foundations of Spacetime Physics: Philosophical Perspectives. Routledge.

Bekenstein, J. D. (1973, Apr). Black holes and entropy. *Phys. Rev. D 7*, 2333–2346.

Bell, J. S. and J. S. Bell (2004). *Speakable and unspeakable in quantum mechanics: Collected papers on quantum philosophy.* Cambridge university press.

Belot, G., J. Earman, and L. Ruetsche (1999a). The hawking information loss paradox: The anatomy of a controversy. *The British Journal for the Philosophy of Science 50*(2), 189–229.

Belot, G., J. Earman, and L. Ruetsche (1999b). The Hawking information loss paradox: the anatomy of controversy. *Brit. J. Phil. Sci. 50*(2), 189–229.

Bena, I., A. Puhm, and B. Vercnocke (2012). Non-extremal black hole microstates: fuzzballs of fire or fuzzballs of fuzz? *Journal of High Energy Physics 2012*(12), 1–34.

Bombelli, L., R. K. Koul, J. Lee, and R. D. Sorkin (1986, Jul). Quantum source of entropy for black holes. *Phys. Rev. D 34*, 373–383.

Brown, A. R., H. Gharibyan, S. Leichenauer, H. W. Lin, S. Nezami, G. Salton, L. Susskind, B. Swingle, and M. Walter (2021). Quantum gravity in the lab: Teleportation by size and traversable wormholes.

Calabrese, P. and J. Cardy (2009, Dec). Entanglement entropy and conformal field theory. *Journal of Physics A: Mathematical and Theoretical 42*(50), 504005.

Cartan, É. (1923). Sur les variétés à connexion affine et la théorie de la relativité généralisée (première partie). In *Annales scientifiques de l'École normale supérieure*, Volume 40, pp. 325–412.

Chen, Y., V. Gorbenko, and J. Maldacena (2021). Bra-ket wormholes in gravitationally prepared states. *Journal of High Energy Physics 2021*(2), 1–61.

Christensen, M. H., J. Hartong, N. A. Obers, and B. Rollier (2014, Mar). Torsional newton-cartan geometry and lifshitz holography. *Phys. Rev. D 89*, 061901.

Cinti, E., A. Corti, and M. Sanchioni (2022). On entanglement as a relation. *European Journal for Philosophy of Science 12*(1), 1–29.

Cinti, E. and M. Sanchioni (2020). Humeanism in light of quantum gravity.

Cinti, E. and M. Sanchioni (2021a). The devil in the (implicit) details. *International Journal of Theoretical Physics*.

Cinti, E. and M. Sanchioni (2021b). The devil in the (implicit) details. *International Journal of Theoretical Physics 60*(9), 3234–3253.

Cinti, E. and M. Sanchioni (202x). Peeking inside the black hole.

Cremonini, S. (2011). The shear viscosity to entropy ratio: a status report. *Modern Physics Letters B 25*(23), 1867–1888.

Crowther, K., N. S. Linnemann, and C. Wüthrich (2021). What we cannot learn from analogue experiments. *Synthese 198*(16), 3701–3726.

Dai, D.-C., D. Minic, D. Stojkovic, and C. Fu (2020, Sep). Testing the ER = EPR conjecture. *Phys. Rev. D 102*, 066004.

Dardashti, R., S. Hartmann, K. Thébault, and E. Winsberg (2019). Hawking radiation and analogue experiments: A bayesian analysis. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 67*, 1–11.

De Haro, S., D. R. Mayerson, and J. N. Butterfield (2016). Conceptual Aspects of Gauge/Gravity Duality. *Found. Phys. 46*(11), 1381–1425.

De Haro, S., J. van Dongen, M. Visser, and J. Butterfield (2020). Conceptual analysis of black hole entropy in string theory. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 69*, 82–111.

Dong, X., D. Harlow, and A. C. Wall (2016, Jul). Reconstruction of bulk operators within the entanglement wedge in gauge-gravity duality. *Phys. Rev. Lett. 117*, 021601.

Earman, J. and G. Valente (2014). Relativistic causality in algebraic quantum field theory. *International Studies in the Philosophy of Science 28*(1), 1–48.

Einstein, A. (1905). Zur elektrodynamik bewegter körper. *Annalen der physik 4*.

Einstein, A. (1915). Die feldgleichungen der gravitation. *Sitzungsberichte der Königlich Preußischen Akademie der Wissenschaften (Berlin*, 844–847.

Einstein, A. (2013). *Relativity*. Routledge.

Einstein, A., B. Podolsky, and N. Rosen (1935, May). Can quantum-mechanical description of physical reality be considered complete? *Phys. Rev. 47*, 777–780.

Einstein, A. and N. Rosen (1935, Jul). The particle problem in the general theory of relativity. *Phys. Rev. 48*, 73–77.

Esfeld, M. (2004). Quantum entanglement and a metaphysics of relations. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 35*(4), 601–617.

Faulkner, T., A. Lewkowycz, and J. Maldacena (2013). Quantum corrections to holographic entanglement entropy. *Journal of High Energy Physics 2013*(11), 74.

Fletcher, S. C., J. Manchak, M. D. Schneider, and J. O. Weatherall (2018, Aug). Would two dimensions be world enough for spacetime? *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 63*, 100–113.

French, S. (2014). *The structure of the world: Metaphysics and representation*. Oxford University Press.

Gan, W.-C. and F.-W. Shu (2017, oct). Holography as deep learning. *International Journal of Modern Physics D 26*(12), 1743020.

Geng, H., A. Karch, C. Perez-Pardavila, S. Raju, L. Randall, M. Riojas, and S. Shashi (2021). Inconsistency of islands in theories with long-range gravity. *arXiv preprint arXiv:2107.03390*.

Gibbons, G. and N. Warner (2013). Global structure of five-dimensional fuzzballs. *Classical and Quantum Gravity 31*(2), 025016.

Giddings, S. B. (2013a). Nonviolent information transfer from black holes: A field theory parametrization. *Phys. Rev. D 88*(2), 024018.

Giddings, S. B. (2013b). Nonviolent nonlocality. *Phys. Rev. D 88*, 064023.

Giddings, S. B. (2013c). Statistical physics of black holes as quantum-mechanical systems. *Phys. Rev. D 88*, 104013.

Giddings, S. B. (2014). Possible observational windows for quantum effects from black holes. *Phys. Rev. D 90*(12), 124033.

Gooth, J., A. C. Niemann, T. Meng, A. G. Grushin, K. Landsteiner, B. Gotsmann, F. Menges, M. Schmidt, C. Shekhar, V. Süß, et al. (2017). Experimental signatures of the mixed axial–gravitational anomaly in the weyl semimetal nbp. *Nature 547*(7663), 324–327.

Gryb, S., P. Palacios, and K. P. Thébault (2021). On the universality of hawking radiation. *The British Journal for the Philosophy of Science*.

Gürsoy, U. (2016). Improved holographic QCD and the quark–gluon plasma. *Acta Physica Polonica B 47*(12), 2509.

Haag, R. (1996). *Local quantum physics: Fields, particles, algebras.* Springer Science & Business Media.

Halvorson, H. (2007). Algebraic quantum field theory. In J. Earman and J. Butterfield (Eds.), *Handbook of Philosophy of Physics*, pp. 731–922. Elsevier.

Hansen, D., J. Hartong, and N. A. Obers (2019). Gravity between newton and einstein. *International Journal of Modern Physics D 28*(14), 1944010.

Hansen, D., J. Hartong, and N. A. Obers (2020). Non-relativistic gravity and its coupling to matter. *Journal of High Energy Physics 2020*(6), 1–100.

Harlow, D. (2016). Jerusalem Lectures on Black Holes and Quantum Information. *Rev. Mod. Phys. 88*, 015002.

Harlow, D. (2017). The ryu–takayanagi formula from quantum error correction. *Communications in Mathematical Physics 354*(3), 865–912.

Harlow, D. (2018). TASI Lectures on the Emergence of Bulk Physics in AdS/CFT. *PoS TASI2017*, 002.

Hartman, T., Y. Jiang, and E. Shaghoulian (2020). Islands in cosmology. *Journal of High Energy Physics 2020*(11), 1–56.

Hartnoll, S. A. (2009, oct). Lectures on holographic methods for condensed matter physics. *Classical and Quantum Gravity 26*(22), 224002.

Hartong, J. and N. A. Obers (2015). Hořava-lifshitz gravity from dynamical newton-cartan geometry. *Journal of High Energy Physics 2015*(7), 1–44.

Hartong, J., N. A. Obers, and M. Sanchioni (2016). Lifshitz Hydrodynamics from Lifshitz Black Branes with Linear Momentum. *JHEP 10*, 120.

Hawking, S. W. (1975). Particle creation by black holes. *Comm. Math. Phys. 43*(3), 199–220.

Hawking, S. W. (1976, Nov). Breakdown of predictability in gravitational collapse. *Phys. Rev. D 14*, 2460–2473.

Hayden, P. and G. Penington (2019a). Learning the Alpha-bits of Black Holes. *JHEP 12*, 007.

Hayden, P. and G. Penington (2019b, Dec). Learning the alpha-bits of black holes. *Journal of High Energy Physics 2019*(12).

Hubeny, V. E., M. Rangamani, and T. Takayanagi (2007, Jul). A covariant holographic entanglement entropy proposal. *Journal of High Energy Physics 2007*(07), 062–062.

Huggett, N. and C. Wüthrich (2013, aug). Emergent spacetime and empirical (in)coherence. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 44*(3), 276–285.

Jaffe, A. and E. Witten (2006). Quantum yang-mills theory. *The millennium prize problems 1*, 129.

Jafferis, D., A. Zlokapa, J. D. Lykken, D. K. Kolchmeyer, S. I. Davis, N. Lauk, H. Neven, and M. Spiropulu (2022). Traversable wormhole dynamics on a quantum processor. *Nature 612*(7938), 51–55.

Jafferis, D. L., A. Lewkowycz, J. Maldacena, and S. J. Suh (2016). Relative entropy equals bulk relative entropy. *JHEP 06*, 004.

Jahnke, V. (2019). Recent developments in the holographic description of quantum chaos.

Jaksland, R. (2018). Probing spacetime with a holographic relation between spacetime and entanglement. *http://philsci-archive.pitt.edu/15415/*.

Jaksland, R. (2020). Entanglement as the world-making relation: Distance from entanglement.

James Read, Marco Sanchioni, W. W. (202x). Underdetermination in general relativity.

Jarrett, J. P. (1984). On the physical significance of the locality conditions in the bell arguments. *Noûs 18*(4), 569–589.

Johansson, L.-G. and K. Matsubara (2011). String theory and general methodology: A mutual evaluation. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 42*(3), 199–210.

Kang, M. J. and D. K. Kolchmeyer (2021). Entanglement wedge reconstruction of infinite-dimensional von neumann algebras using tensor networks. *Physical Review D 103*(12), 126018.

Ladyman, J. (1998). What is structural realism? *Studies in History and Philosophy of Science Part A 29*(3), 409–424.

Ladyman, J., D. Ross, J. Collier, D. Spurrett, D. Spurrett, J. G. Collier, et al. (2007). *Every thing must go: Metaphysics naturalized*. Oxford University Press on Demand.

Lakatos, I. (1976). *Falsification and the methodology of scientific research programmes*. Springer.

Landsteiner, K., E. Megias, and F. Pena-Benitez (2011). Gravitational anomaly and transport phenomena. *Physical review letters 107*(2), 021601.

Le Bihan, B. (2020). String theory, loop quantum gravity and eternalism. *European Journal for Philosophy of Science 10*(2), 1–22.

Le Bihan, B. and J. Read (2018). Duality and ontology. *Philosophy Compass 13*(12), e12555.

Leonard, H. S. and N. Goodman (1940). The calculus of individuals and its uses. *Journal of Symbolic Logic 5*(2), 45–55.

Lewis, D. (1974). Causation. *The journal of philosophy 70*(17), 556–567.

Lewis, D. (2013). *Counterfactuals*. John Wiley & Sons.

Maldacena, J. (1999a). The large n limit of superconformal field theories and supergravity. *International Journal of Theoretical Physics 38*(4), 1113–1133.

Maldacena, J. (1999b). The large-n limit of superconformal field theories and supergravity. *International journal of theoretical physics 38*(4), 1113–1133.

Maldacena, J. (2003). Eternal black holes in anti-de sitter. *Journal of High Energy Physics 2003*(04), 021.

Maldacena, J. and L. Maoz (2004). Wormholes in ads. *Journal of High Energy Physics 2004*(02), 053.

Maldacena, J. and L. Susskind (2013). Cool horizons for entangled black holes. *Fortsch. Phys. 61*, 781–811.

Maldacena, J. M. (1999c). The Large N limit of superconformal field theories and supergravity. *Int. J. Theor. Phys. 38*, 1113–1133.

Marolf, D. and J. Polchinski (2013). Gauge-gravity duality and the black hole interior. *Physical review letters 111*(17), 171301.

Marolf, D. and J. Polchinski (2016). Violations of the Born rule in cool state-dependent horizons. *JHEP 01*, 008.

Matarese, V. (2019). Loop Quantum Gravity: A New Threat to Humeanism? Part I: The Problem of Spacetime. *Found. Phys. 49*(3), 232–259.

Mathur, S. D. (2009). The information paradox: a pedagogical introduction. *Classical and Quantum Gravity 26*(22), 224001.

Mathur, S. D. and D. Turton (2014a). Comments on black holes i: the possibility of complementarity. *Journal of High Energy Physics 2014*(1), 1–32.

Mathur, S. D. and D. Turton (2014b). The flaw in the firewall argument. *Nuclear Physics B 884*, 566–611.

Maudlin, T. (2002). *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics*. Blackwell.

Maudlin, T. (2011). *Quantum non-locality and relativity: Metaphysical intimations of modern physics*. John Wiley & Sons.

Maudlin, T. (2017a). (Information) Paradox Lost. last accessed 16-09-2020.

Maudlin, T. (2017b). (information) paradox lost. *arXiv preprint arXiv:1705.03541*.

McKenzie, K. (2017). Ontic structural realism. *Philosophy Compass 12*(4), e12399.

Menzies, P. and H. Beebee (2020). Counterfactual Theories of Causation. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Winter 2020 ed.). Metaphysics Research Lab, Stanford University.

Michelson, A. A. and E. W. Morley (1887). On the relative motion of the earth and of the luminiferous ether. *Sidereal Messenger, vol. 6, pp. 306-310 6*, 306–310.

Morganti, M. (2009). A new look at relational holism in quantum mechanics. *Philosophy of Science 76*(5), 1027–1038.

Mukohyama, S. (2010, Nov). Hořava–lifshitz cosmology: a review. *Classical and Quantum Gravity 27*(22), 223101.

Mussardo, G. (2010). *Statistical field theory: an introduction to exactly solved models in statistical physics*. Oxford University Press.

Myrvold, W. C. (2015). Lessons of bell's theorem: Nonlocality, yes; action at a distance, not necessarily.

Nezami, S., H. W. Lin, A. R. Brown, H. Gharibyan, S. Leichenauer, G. Salton, L. Susskind, B. Swingle, and M. Walter (2021). Quantum gravity in the lab: Teleportation by size and traversable wormholes, part ii.

Oliver, A. and T. Smiley (2004). Multigrade predicates. *Mind 113*(452), 609–681.

Olson, S. and T. C. Ralph (2012). Extraction of timelike entanglement from the quantum vacuum. *Phys. Rev. A 85*, 012306.

Page, D. N. (1993, Dec). Information in black hole radiation. *Physical Review Letters 71*(23), 3743–3746.

Papadodimas, K. and S. Raju (2013). An Infalling Observer in AdS/CFT. *JHEP 10*, 212.

Papadodimas, K. and S. Raju (2016). Remarks on the necessity and implications of state-dependence in the black hole interior. *Phys. Rev. D 93*(8), 084049.

Penington, G. (2019). Entanglement wedge reconstruction and the information paradox.

Penington, G. (2020). Entanglement wedge reconstruction and the information paradox. *Journal of High Energy Physics 2020*(9), 1–84.

Penington, G., S. H. Shenker, D. Stanford, and Z. Yang (2019). Replica wormholes and the black hole interior.

Penington, G., S. H. Shenker, D. Stanford, and Z. Yang (2020). Replica wormholes and the black hole interior.

Penington, G., S. H. Shenker, D. Stanford, and Z. Yang (2022). Replica wormholes and the black hole interior. *Journal of High Energy Physics 2022*(3), 1–87.

Penrose, R. (2007). *The Road to Reality: A Complete Guide to the Laws of the Universe.* Vintage Series. Vintage Books.

Policastro, G., D. T. Son, and A. O. Starinets (2001). Shear viscosity of strongly coupled n= 4 supersymmetric yang-mills plasma. *Physical Review Letters 87*(8), 081601.

Preskill, J. (2015). *Lecture Notes for Physics 229:Quantum Information and Computation.* CreateSpace Independent Publishing Platform.

Read, J. and N. J. Teh (2022). Newtonian equivalence principles. *Erkenntnis*, 1–25.

Redhead, M. (1987). *Incompleteness, Nonlocality, and Realism: A Prolegomenon to the Philosophy of Quantum Mechanics.* Oxford University Press.

Rovelli, C. (2016). An argument against the realistic interpretation of the wave function. *Foundations of Physics 46*(10), 1229–1237.

Rovelli, C. (2018). Space is blue and birds fly through it. *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences 376*(2123), 20170312.

Ryu, S. and T. Takayanagi (2006, May). Holographic derivation of entanglement entropy from the anti–de sitter space/conformal field theory correspondence. *Phys. Rev. Lett. 96*, 181602.

Saad, P., S. H. Shenker, and D. Stanford (2019). Jt gravity as a matrix integral.

Sachdev, S. and J. Ye (1993). Gapless spin-fluid ground state in a random quantum heisenberg magnet. *Physical review letters 70*(21), 3339.

Sainsbury, R. M. (1995). Vagueness, ignorance, and margin for error. *The British Journal for the Philosophy of Science 46*(4), 589–601.

Sainsbury, R. M. (2009). *Paradoxes.* Cambridge University Press.

Sanchioni, M. (202x). Lakatos rational reconstruction and quantum gravity.

Schäfer, T. and D. Teaney (2009). Nearly perfect fluidity: from cold atomic gases to hot quark gluon plasmas. *Reports on Progress in Physics 72*(12), 126001.

Schaffer, J. (2016). Grounding in the image of causation. *Philosophical Studies 173*(1), 49–100.

Shen, C., U. Heinz, P. Huovinen, and H. Song (2011). Radial and elliptic flow in pb+ pb collisions at energies available at the cern large hadron collider from viscous hydrodynamics. *Physical Review C 84*(4), 044903.

Smolin, L. (2007). *The trouble with physics: the rise of string theory, the fall of a science, and what comes next.* HMH.

Son, D. T. (2013, 6). Newton-Cartan Geometry and the Quantum Hall Effect.

Sorensen, R. (2003). *A Brief History of the Paradox: Philosophy and the Labyrinths of the Mind.* Oxford University Press.

Starr, W. (2021). Counterfactuals. In E. N. Zalta (Ed.), *The Stanford Encyclopedia of Philosophy* (Summer 2021 ed.). Metaphysics Research Lab, Stanford University.

Strominger, A. and C. Vafa (1996, Jun). Microscopic origin of the bekenstein-hawking entropy. *Physics Letters B 379*(1-4), 99–104.

Susskind, L. (2008). *The Black Hole War: My Battle with Stephen Hawking to Make the World Safe for Quantum Mechanics*. Little, Brown.

Susskind, L. (2016a). Copenhagen vs everett, teleportation, and er= epr. *Fortschritte der Physik 64*(6-7), 551–564.

Susskind, L. (2016b). Copenhagen vs Everett, Teleportation, and ER=EPR. *Fortsch. Phys. 64*(6-7), 551–564.

Susskind, L., L. Thorlacius, and J. Uglum (1993, Oct). The stretched horizon and black hole complementarity. *Physical Review D 48*(8), 3743–3761.

Teller, P. (1986). Relational holism and quantum mechanics. *The British Journal for the Philosophy of Science 37*(1), 71–81.

Unruh, W. G. (1981). Experimental black-hole evaporation? *Physical Review Letters 46*(21), 1351.

Unruh, W. G. and R. M. Wald (1995a). Evolution laws taking pure states to mixed states in quantum field theory. *Physical Review D 52*(4), 2176.

Unruh, W. G. and R. M. Wald (1995b). On evolution laws taking pure states to mixed states in quantum field theory. *Phys. Rev. D 52*, 2176–2182.

Unruh, W. G. and R. M. Wald (2017). Information Loss. *Rept. Prog. Phys. 80*(9), 092002.

Visser, M. (1998). Acoustic black holes: horizons, ergospheres and hawking radiation. *Classical and Quantum Gravity 15*(6), 1767.

Vistarini, T. (2019). *The Emergence of Spacetime in String Theory*. Routledge.

Wallace, D. (2012). *The emergent multiverse: Quantum theory according to the Everett interpretation*. Oxford University Press.

Wallace, D. (2018). The case for black hole thermodynamics part i: Phenomenological thermodynamics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 64*, 52 – 67.

Wallace, D. (2019). The case for black hole thermodynamics part ii: Statistical mechanics. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics 66*, 103 – 117.

Wallace, D. (2020a). Why black hole information loss is paradoxical. In N. Huggett, K. Matsubara, and C. Wüthrich (Eds.), *Beyond Spacetime: The Foundations of Quantum Gravity*, pp. 209–236. Cambridge University Press.

Wallace, D. (2020b). Why Black Hole Information Loss is Paradoxical. In N. Huggett, K. Matsubara, and C. Wüthrich (Eds.), *Beyond Spacetime*. Cambridge University Press.

Wallace, D. (2022). Quantum gravity at low energies. *Studies in History and Philosophy of Science 94*, 31–46.

Weinberg, S. and F. Wilczek (1993). Dreams of a final theory. *Physics Today 46*(4), 59.

Weinstein, G. (2023). A comment on "traversable wormhole dynamics on a quantum processor".

Williams, P. (2018, January). Renormalization group methods. To appear in the Routledge Companion to Philosophy of Physics, Eleanor Knox and Alastair Wilson (eds).

Witten, E. (1998). Anti de sitter space and holography. *arXiv preprint hep-th/9802150*.

Witten, E. (2020, Mar). A mini-introduction to information theory. *La Rivista del Nuovo Cimento 43*(4), 187–227.

Wuthrich, C. (2017). Are black holes about information?

Wüthrich, C. (2019, April). When the actual world is not even possible. 20 pages, 1 figure; April 2019 revision; for George Darby, David Glick, and Anna Marmodoro (eds.), The Foundation of Reality: Fundamentality, Space and Time, Oxford University Press.