What Can we Learn (and not Learn) from Thought Experiments in Black Hole Thermodynamics?

Rawad El Skaf
Department of Mathematics, Politecnico di Milano
Patricia Palacios
Department of Philosophy, University of Salzburg

Abstract

Scientists investigating the thermal properties of black holes rely heavily on theoretical and non-empirical tools, such as mathematical derivations, analogue experiments and thought experiments. Although the use of mathematical derivations and analogue experiments in the context of black hole physics has recently received a great deal of attention among philosophers of science, the use of thought experiments in that context has been almost completely neglected. In this paper, we will start filling this gap by systematically analyzing the epistemic role of the two thought experiments that gave birth to black hole thermodynamics, namely Wheeler's demon and Geroch's engine. We will argue that the two main epistemic functions of these thought experiments are to reveal and resolve inconsistencies, in line with El Skaf's (2021) approach to TEs. We will, then, go beyond El Skaf's approach by stressing an important difference between the strategies employed to assess the reliability of each epistemic function.

1 Introduction

Black hole thermodynamics (BHT) is a discipline that combines theoretical statements coming from three main theories: quantum mechanics, general relativity and thermodynamics. Although BHT has attracted a great deal of attention in the last decades, it still lacks direct empirical support, which is not surprising, given that black holes are experimentally inaccessible, barely observable and surely unmanipulable systems. In this context, thought experiments (TEs) instead of empirical (real or laboratory) experiments have proved to be one of the most important tools for getting novel insights about the thermal properties of black holes. But to what extent can we trust the results coming from TEs in BHT? And what are the limitations of the knowledge that can be obtained on the basis of TEs?

Some physicists (e.g. Susskind 2008, Polchinski 2017) have stressed the importance of TEs for bringing to light paradoxes between fundamental theories in the context of black holes, and some philosophers of science have even suggested that black hole TEs can give some theoretical support to the idea that black holes have thermodynamic properties. Curiel (2014), for instance, says: "Why assume a classical black hole has an entropy in the first place? The best answer to this is implicit in the series of thought-experiments". (p. 16) Similarly, Wüthrich (2019) argues: "Gedanken experiments concerning the limits of the amount of thermodynamic work that can or cannot be extracted from black holes lend some support to the idea that black holes are thermodynamic in nature." (p. 221).

However, despite the essential role that TEs seem to play in BHT, there has been surprisingly little philosophical work on this topic. In fact, neither philosophers of science working on the epistemology of (scientific) TEs nor philosophers of physics working on BHT have carried out a systematic analysis on the use of TEs in BHT yet[1]. On the one hand, philosophers of science working on the epistemology of TEs (e.g. Norton 1991, 1996, 2004, Brown 1991, Nersessian 1992, Bokulich 2001, Stuart 2018, El Skaf 2021) have mainly focused on case studies taken from the history of physics, from Galileo to Einstein's TEs. On the other hand, most of the philosophical work around BHT has focused either on the main *calculations* that give theoretical support to the idea that black holes are thermodynamic objects (e.g. Dougherty and Callender 2016, Wallace 2018, 2019, Belot, Earman, and Ruetsche 1999 and Earman 2011, Gryb et al. 2021) or on the use of *analogue experiments*, which are real experiments performed in systems different but analogous to black holes (e.g. Crowther et al., 2021; Dardashti et al., 2017, 2019).

We believe that this lack of philosophical attention on the use of TEs in BHT is unfortunate. First of all, because BHT illustrates, perhaps better than any other discipline, the importance of TEs in ongoing physics and, in this way, it makes an interesting case study for the philosophical analysis around TEs in science. Second, because, by being beyond the reach of direct empirical testing, BHT is an ideal arena to understand the importance of TEs when direct empirical evidence is entirely lacking. In this paper, we will start filling this existing gap by systematically analyzing the role of TEs in BHT. In particular, we will argue that the two main epistemic functions of TEs in black hole physics are to reveal and resolve inconsistencies, in line with what has been defended by El Skaf (2021). We will, then, go beyond El Skaf's approach by stressing an important difference between the strategies employed to assess the reliability of each epistemic function.

This paper is organized as follows. In Section 2, we will introduce Wheeler's TE and Geroch's engine TE. We will argue that Wheeler's TE reveals a tension between general relativity's no-hair theorem and the second law of thermodynamics, and we will then explain how Bekenstein (1972) attempts to resolve

[1] An exception is the paper by Weinstein (2021), which focuses on the use TEs in black hole physics, but rather from a historical point of view.

this tension by formulating the Generalised Second Law (GSL). We will, then, introduce Geroch's TE and argue that it reveals an inconsistency between other statements of general relativity associated with the existence of an event horizon and both the second law of thermodynamics and GSL. In the same section, we will discuss different proposals suggested in the literature to solve this inconsistency, including Bekenstein's entropy bound (Bekenstein 1981) and Unruh and Wald's buoyancy effect (Unruh and Wald 1982). In Section 3, we will review the philosophical literature on TEs with special focus on El Skaf (2021)'s account, which stresses that the main functions of some TEs are to reveal and resolve inconsistencies. In Section 4, we will re-evaluate Wheeler and Geroch TEs and will argue that their main epistemic functions are to unveil well-hidden external inconsistencies and to suggest possible ways to resolve them. Here, we will extend El Skaf's approach by pointing out that the justification of these two epistemic functions substantially differ. After that, we will briefly discuss other TEs used in black hole physics associated with the so-called "Information Loss Paradox" and review some of the proposed resolutions. Finally, in Section 4.4, we will discuss other theoretical tools that may play a role in the acceptance or rejection of a given resolution, such as analogue experiments and mathematical derivations.

2 TEs in Black Hole Physics: Wheeler and Geroch

TEs are widely used in investigating BHT. As we have already noted, this is unsurprising, given the nature of their object of inquiry. In this section, we will introduce the TEs that initiated the field of BHT, namely Wheeler's TE and Geroch's Engine TE. This will serve as a starting point for a more profound analysis around these TEs, which will be carried out in Section 4.

2.1 Wheeler's Demon and the Generalized Second Law

In a paper of 1980, Jacob Bekenstein recounts a discussion he had with John Wheeler while writing his doctoral dissertation (Bekenstein 1980, p. 24). During this discussion, Wheeler suggested to Bekenstein to consider the following situation: Two cups of tea at different temperatures are brought into thermal contact. After a while they will equilibrate into a common temperature. One should, then, imagine that a black hole is passing in front of them and that one throws the two cups into it (see Figure 1). What happens then? A few months later, Bekenstein came up with an answer in his celebrated paper "Black Holes and the Second Law" (Bekenstein 1972), which is one of the papers that gave birth to the field of BHT. In this paper, he reformulates Wheeler's TE in the following way:

> Let an observer drop or lower a package of entropy into a black hole; the entropy of the exterior world decreases. Furthermore, from an

> exterior observer's point of view a black hole in equilibrium has only three degrees of freedom: mass, charge and angular momentum [...]. Thus, once the black hole has settled down to equilibrium, there is no way for the observer to determine its interior entropy. Therefore, he cannot exclude the possibility that the total entropy of the universe may have decreased in the process. It is in this sense that the second law appears to be transcended. (Bekenstein 1972, p. 737)

One can see that instead of bringing two cups of tea together and then throwing them into a black hole, Bekenstein simply imagined that an observer drops a "package of entropy into a black hole". Furthermore, in the description of this experimental set-up, or what we call here "scenario", he explicitly includes a theoretical statement of general relativity (GR) known under the name of "no-hair theorem", which states that black holes are uniquely characterized by three free parameters: mass, angular momentum, and the electric charge.[2] This means that, according to the no-hair theorem of GR, black holes are extremely simple objects, so simple that an external observer will not be able to distinguish between black holes made from disparate kinds of matter, if they have the same mass, charge and angular momentum (Ruffini and Wheeler 1971). A consequence of this is that an external observer will not be able to measure or observe any other property of a system that is "thrown" into a black hole, including the entropy of a cup of tea. With that in mind, we can understand the difficulty raised by Wheeler's TE: Once the package of entropy is thrown into a black hole, the no-hair theorem states that one cannot rule out the possibility that the total entropy of the universe may have decreased in the process. This is so, because an outside observer can no longer determine its inner entropy.

[2] The No-Hair Theorem comes from a remarkable series of results, collectively known under the name of "no-hair theorem."

Figure 1 Illustration of Weehler's original TE. In this experiment, it is imagined that two cups of tea are dropped into a black hole

For Bekenstein (1972, 1980), Wheeler's TE tries to show that the second law can be "transcended", which means for him that it loses its predictive power or, in other words, that it is observationally meaningless. However, a closer look at Weehler's TE revails that it does not directly lead to the "transcendence" of the second law, but instead that it unveils a tension between some of the theoretical statements that are used to describe the experimental scenario. More precisely, this TE reveals an inconsistency between Wheeler's no-hair theorem and the second law of thermodynamics, which states that the entropy of an isolated system cannot decrease. Bekenstein himself seems to recognize this, when he says:

> [A]s a graduate student of Wheeler's at Princeton I found "black holes have no hair" distressing for a reason he brought home to me in a 1971 conversation. The principle, he argued, allows a wicked creature – call it Wheeler's demon – to commit the perfect crime against the second law of thermodynamics" (Bekenstein 1980, p.24).

Despite this comment, Bekenstein at the time did not see the inconsistency revealed by the TE as a challenge for the validity of the no-hair theorem, but only for the second law. We will come back to the analysis of this inconsistency in Section 4.

Let us now look at how Bekenstein proposes to save the second law. His idea was ingenious and simple: In 1972, Bekenstein (1972) proposed to generalise the second law of thermodynamics, so as to include the entropy of a black hole. More precisely, he proposes that the sum of the change of the black hole entropy dS_{BH} and the common entropy outside the black hole dS_M must never decrease or, in his own words, that "common entropy plus black-hole entropy never decreases" (Bekenstein 1972, p. 738). Formally, this can be written as follows:

$$dS_{total} = dS_{BH} + dS_M \geq 0, \tag{1}$$

which is now known as the *Generalized Second Law of Thermodynamics* (GSL).

Bekenstein (1972) defined the entropy of the black hole as proportional to the surface area A of the event horizon of the black hole:

$$S_{BH} = \frac{\eta k A}{L_p^2}, \tag{2}$$

where L_p is the Planck length: $(\hbar G/c^3)^{1/2}$, k is Boltzmann's constant, and η is a constant number of order unity. The choice of the area of a black hole as a measure of its entropy is motivated by Christodoulou (1970) and Hawking's area theorem (Hawking 1971), which states that the area A of a black hole never decreases:

$$dA \geq 0. \tag{3}$$

In fact, for Bekenstein, the area appeared "to be the only one of [the black hole] properties having this entropylike behavior which is so essential if the second law as we have stated it is to hold when entropy goes down a black hole." (Bekenstein 1972, p. 104).

Note that the black hole entropy (eq. 2) links a thermodynamic quantity (entropy) with a gravitational one (surface area). Furthermore, it also establishes an important connection with quantum mechanics, since this link breaks down in the classical limit $\hbar \to 0$ (Bekenstein 1980). This means that Bekenstein's definition of black hole entropy establishes a deep relation between three main theories: thermodynamics, general relativity and quantum mechanics.

2.2 Geroch's Engine and the Entropy Bound

We have seen that Bekenstein's strategy to exorcise Weehler's demon was to ascribe entropy to black holes. In his 1972 paper, he also discusses another TE that was supposed to show that the second law of thermodynamics may be not only "transcended" but manifestly violated. He describes this TE as follows:

> A method for violating the second law has been proposed by GEROCH: By means of a string one slowly lowers a body of rest mass m and nonzero temperature toward a Schwarzschild black hole of mass M. By the time the body nears the horizon, its energy as measured from infinity, $E = m(1-2M/r)^{\frac{1}{2}}$, is nearly zero; the body has already done work m on the agent which lowers the string. At

this point the body is allowed to radiate into the black hole until its rest mass is $m - \Delta m$. Finally, by expending work $m - \Delta m$, one hauls the body back up.

The net result: a quantity of heat Δm has been completely converted into work. Furthermore, since the addition of the radiation to the black hole takes place at a point where $(1 - 2M/r)^{\frac{1}{2}} \approx 0$, the mass of the black hole is unchanged. Thus the black hole appears to be unchanged after the process. This implies a violation of the second law: "One may not transform heat entirely into work without compensating changes taking place in the surroundings." (Bekenstein 1972, p.373)

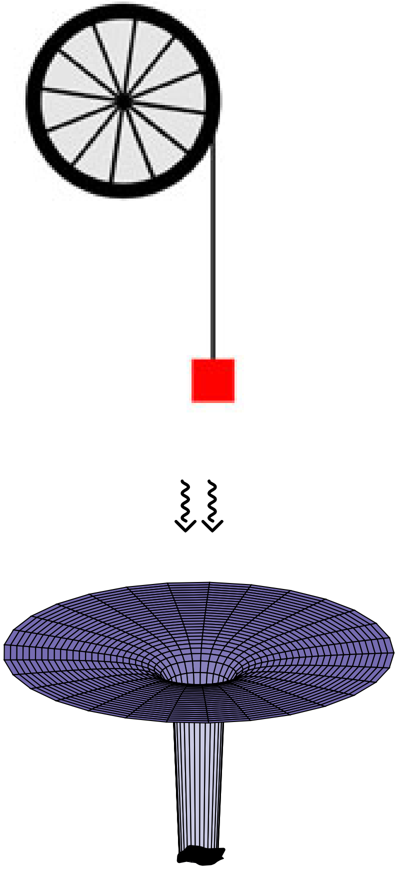

Figure 2 Illustration of Geroch's TE. In this experiment, it is imagined that a box with entropy S is slowly lowered towards the event horizon

In order to understand how this TE works, we need to understand some details and assumptions underlying the scenario of the TE. This TE was mentioned by Geroch in a colloquium at Princeton in December 1971, which was attended by Bekenstein. In this TE (from now on "Geroch's TE"), Geroch asked the audience to consider a heat engine that uses a Schwarzschild black hole as an energy sink (Figure 2). Knowing that the Schwarzschild metric is:

$$ds^2 = (1 - \frac{2R_g}{r})c^2dt^2 - \frac{dr^2}{1 - \frac{2R_g}{r}} - r^2(d\theta^2 sin^2\phi^2), \tag{4}$$

where R_g is the gravitational radius, defined as:

$$R_g = \frac{GM}{c^2}, \tag{5}$$

and M is the mass of the black hole, the experiment consist of the following steps:[3]

1. We fill a box (red box in Figure 2) with heat radiation of energy $E = m(1 - 2M/r)^{1/2}$, temperature T and entropy S. We assume that the box and the rope have no weight. We also assume that the box has perfectly reflecting walls.

2. We slowly winched the box towards the horizon of the Schwarzschild black hole, so that $r \to 2R_g$.

3. Since the total energy of the radiation consists of the heat energy and gravitational energy, as the box descends, the negative gravitational energy grows, thus paying for the positive energy being given to the reservoir. Eventually, the gravitational energy cancels the positive heat energy, so that the total energy E of the body as measured from infinity is nearly zero. In fact, since $r \to 2R_g$, it follows from equation (5) that $E = m(1 - 2M/r)^{1/2}$ goes to zero.

4. We then open the red box and allow radiation to escape to the black hole until its rest mass is $m - \Delta m$.

5. The box can be pulled up back at expense of $m - \Delta m$, which means that the quantity Δm can be completely converted into work.

Since the energy goes to zero, the mass of the black hole must thus remain unchanged in the process, which suggests that the black hole may end up in the same state it began.

Furthermore, as stated in step 5, Geroch's TE was used to show that a black hole can run a Carnot cycle with a hundred percent efficiency. A consequence of this is that the temperature of the black hole T_{BH} must be zero. In fact, the Carnot efficiency η of the heat engine is:

[3] Unless specified otherwise, we set $G = c = k = \hbar = 1$.

$$\eta \leq (1 - \frac{T_{BH}}{T_{Ra}}), \tag{6}$$

where T_{Ra} is the temperature of the radiation coming from the box and T_{BH} is the temperature of the black hole. If all the energy is converted into work, then it follows that the efficiency is 1 and, consequently, $T_{BH} = 0$. Geroch, in fact, used this argument to stress that black holes are systems at zero temperature (Weinstein 2021; Wald 2020). However, if black holes were in fact systems at zero temperature, this would imply not only a violation of the second law, but also of GSL, because this law assigns a finite non-zero entropy to the black hole, which, according to the first law of black hole thermodynamics, also requires attributing a finite non-zero temperature to black holes (we will come back to the analysis of this inconsistency in Section 4).[4]

According to Wald (2020), Bekenstein was concerned about these results, precisely because they appeared to contradict GSL.

> It seems clear that Bekenstein must have immediately realized that assigning an absolute zero physical temperature to a black hole would lead to severe consistency problems with black hole thermodynamics. In particular, Geroch's suggestion of lowering a box of matter containing entropy all the way to the horizon of a black hole could certainly be used to violate any proposal for a generalized second law, since, in this process, entropy would be lost, but the black hole would end up in the same state in which it began. (Wald 2020, p. 6)

Hawking's (1974) prediction that black holes emit radiation with temperature $T = \kappa/2\pi$, that is, proportional to its surface gravity κ, gave important support to BHT, but the problem raised by Geroch's TE remained (Wald 2020). Indeed, if one could lower a box arbitrarily close to the horizon, the entropy of the box could still escape to the black hole without increasing the black hole area. This would be in contradiction with GSL, because the entropy in the exterior of the black hole would decrease without an increase in the entropy of the black hole, $dS_{BH} = 0$, which is associated with its area. This means: $dS_{total} = dS_M < 0$.

The latter motivated Bekenstein to find a resolution for the TE that salvaged GSL. He had the intuition that in order to violate GSL by lowering a box towards a black hole, the box would have to be extremely close to the horizon before dropping radiation in and he doubted that this was physically possible (Wald 2020). In his 1981, he proposes a resolution that became known as "Bekenstein bound"[5]:

[4] The general form of the first law of black hole thermodynamics takes the form: $\delta M = \frac{k}{2\pi}\delta S_{BH} + \delta J + ...$, where "..." denote possible additional contributions coming from long range matter field and $S_{BH} = A/4$ (see Wald 2001 for details).

[5] The basic idea of a physical bound was already present in earlier papers (e.g. Bekenstein 1973, 1974)

> In fact, black-hole physics yields a specific form for the upper bound on S/E for systems with negligible self-gravity. According to the generalized second law of thermodynamics, the sum of the thermal entropy outside a black hole and the black hole entropy (1/4 of the horizon's surface area) should never decrease. Now, it has long being known that when a stationary hole absorbs a body with negligible self-gravity, energy E and effective radius R (...), the hole's surface area must increase by at least $8\pi ER$. Since one can arrange the absorption process so that this minimal increase can be attained, the second law will be violated unless the body's entropy (what disappears from the hole's exterior) cannot exceed $2\pi ER$. Thus we obtain the bound on S/E to weakly gravitating bodies (Bekenstein 1981, p. 288).

The basic idea was, then, to impose a physical bound that cannot be exceed by the box or any other physical system. The Bekenstein bound is:

$$S/E \leq 2\pi R, \tag{7}$$

where S is the entropy, E is the energy, and R is the effective (or "circumscribing") radius of the body, when the radiation is dropped into the black hole. This bound is derived from an equation that determines the mass increase of the black hole when radiation is dropped into it (see Bekenstein 1981 for details). Bekenstein, then, stresses that if S does not exceed $2\pi ER$, then GSL would not be violated. However, in the following subsection, we will discuss some objections to this resolution.

2.3 Other Resolutions of Geroch's TE

As Robert Wald (2020) recalls it, he and Bill Unruh were unhappy with Bekenstein 1981's resolution of Geroch's TE for two main reasons: (i) The bound didn't appear to be sufficiently general and robust to avoid a violation of the generalized second law. In fact, they point out that if one uses, for instance, a rectangular box instead of a square box, it would be necessary for the quantity "R" in the bound to be the shortest dimension of the box, whereas the arguments in favor of the bound took R to be the largest dimension. They also point out that if one imagines increasing the number of species n of massless particles in nature, then one could make the S/E ratio arbitrarily large for a given R, thus violating Bekenstein bound (Unruh and Wald 1982). (ii) They also thought that the consistency of black hole thermodynamics should not depend on some property of matter that would not otherwise be needed for the consistency of thermodynamics. In other words, they took Bekenstein's solution to be *ad hoc*.

Motivated by these concerns, Unruh and Wald (1982) came up with a different resolution of Geroch's TE that can be summarized as follows. They first noted that quantum effects, like Hawking radiation, are very small for large black holes, but they become important for quasi-stationary bodies near the horizon, such as the case of a box of energy E and entropy S being slowly lowered towards the horizon. In fact, they showed that these bodies would undergo an enormous acceleration and therefore feel the effects of the quantum "thermal atmosphere" surrounding the black hole. They then argued that the temperature gradient in this thermal atmosphere will produce a pressure gradient and, therefore, a *buoyancy force* on the box, which becomes infinitely large in the limit as the box is lowered towards the horizon. The result is that this buoyancy force will prevent Geroch's box from reaching the horizon. In fact, the optimal place from which to drop a box of matter into the black hole will no longer be the horizon but rather the "floating point" of the box, which corresponds to the point in which the weight of the box is equal to the weight of the displaced thermal atmosphere. Finally, they showed that the minimum area increase of the black hole when dropping the matter into it from the floating point is no longer zero, but the amount just sufficient to prevent a violation of GSL.

However, the discussion did not end there. In a series of papers, Bekenstein (e.g. 1983, 1994, 1999) criticized Unruh and Wald's resolution, by pointing out potential deficiencies in their analysis. In 1994, for instance, he showed that under certain assumptions concerning the size of the box and the location of the floating point, the buoyancy force of the thermal atmosphere can be shown to be zero, which means that this resolution cannot assure the validity of GSL for all cases. In 1999, Bekenstein showed that under other conditions, the box size at the floating point can be smaller than the typical wavelengths in the thermal atmosphere, which can likely decrease the magnitude of the buoyancy force. Unruh and Wald responded to these and other criticisms in a series of papers (Unruh and Wald 1983, Pelath and Wald 1999). However, they never reached a consensus and the question of whether the appeal to the buoyancy force is the best strategy to resolve the contradictions posed by Geroch's TE remained open (Page 2020, Wald 2020).[6]

More recently, some physicists have suggested an alternative entropy bound, namely: $S \leq A/4$, which is associated to the "holographic principle" that roughly states that the physics in every spatial region can be described in terms of the degrees of freedom associated with the boundary of the region ('t Hooft 1988, Susskind 1995). This bound has the advantage that it does not make reference to E and so it avoids problems associated with defining E in curved spacetime. However, like Bekenstein's bound, it may fail for physically reasonable systems (Wald 2001). We will come back to the discussion on the robustness of the proposed resolutions in Section 4.2.

3 On the Epistemology of TEs in the History of Science

In the two case studies examined in Section 2, it appears that physicists have arrived at important results in BHT by reasoning through TEs. In fact, in-

[6] Jacob Bekenstein passed away in August 2015.

stead of conducting direct empirical (real world or laboratory) experiments, which could potentially provide new empirical data, physicists have based their discussion on merely imagined experimental set-ups. For empirically minded philosophers of science, this is extremely puzzling, since it appears that scientists have gained some new insight about the physical world, in this case about black holes, without conducting any direct empirical experiment.

In the philosophy of science literature, philosophers have tried to explain such "epistemic magic" (Norton, 2004), by focusing on other TEs, mostly from the history of science, such as Galileo's falling bodies TE (Gendler 1998; El Skaf 2018; Palmieri 2005) and several TEs suggested by Einstein (El Skaf 2021; Norton 1991). In this section, we will briefly review this literature with special focus on a recent account defended by El Skaf (2021), which, as we will argue in Section 4, can help us identify and understand the most important epistemic functions of TEs in BHT.

3.1 What Can We Learn From TEs and How?

In the discussion around TEs, philosophers of science have tried to answer the following two interrelated questions. The first is what *kind* of new insight do TEs provide. In other words, what is their epistemic function. The second is *how* can TEs lead to this identified new insight, and that without any new empirical data. Unsurprisingly, philosophers of science have given different answers to the first question. For instance, Norton (e.g. 1991, 1996, 2004) has argued that TEs can always be reconstructed as deductive or inductive arguments. This means that the new insight that TEs provide depends on the type of argument that can be reconstructed on the basis of a TE. If the argument constructed from a TE is deductive, the TE would just serve to rearrange our existing knowledge without adding new content to our web of beliefs. If the argument is inductive, the TE could generalize our knowledge, in the same way as inductive arguments do.

Brown (1991) has defended a different approach. In contrast to Norton, he does not identify TEs with arguments and provides a detailed taxonomy of the different types of TEs, which are associated with different epistemic functions of TEs, such as constructive, conjectural and "platonic". The most interesting type are platonic TEs, which, according to Brown, can provide us with *a priori* access to the laws of nature, and this without *any* new empirical data.[7] In contrast to Brown, Bokulich (2001) has defended that TEs test the *non-empirical virtues* of our theories, such as consistency and explanatory power. More recently, Stuart (2018) has argued that TEs provide us with understanding, not knowledge.

The second question, namely how can TEs generate new insight without any new empirical data, has attracted much attention in philosophy. For instance, in Norton's view (1991, 1996, 2004), TEs are just arguments and, therefore, the conclusion obtained on the basis of TEs is justified in the same way as

[7] Brown identifies Galileo's falling bodies and EPR as instances of platonic TEs (see Brown 1991 for more details and El Skaf 2018, 2021 for criticism).

the conclusion of inductive or deductive arguments. In addition, Norton (1991) contends that the particular experimental details of the imagined experimental arrangement are *irrelevant* and, thus, *eliminable* from the final reconstructed argument. This has been called "Norton's Elimination Thesis" (see Gendler 1998; Brendel 2018; El Skaf 2021).

Contrary to Norton, defenders of the so-called "mental model" account of TEs (e.g. Nersessian 1992, 2007; Miščević 1992) have criticized the idea that TEs are just arguments and they have rejected the view that the justificatory power of TEs can be reduced to the logical structure of their propositional content, and that the experimental details are irrelevant and eliminable. Instead, these accounts, albeit different on their definition of what a mental model is, share the idea that the imagined experimental arrangement of a TE is an essential vehicle that enables us to construct and reason on non-propositional mental models. Nersessian (1992, 2007), for instance, argues that it is the representation relation (usually a structural similarity) between the imagined system and the real world phenomena what does the justificatory work. According to this view, we acquire new knowledge about the real world target system by mentally modelling a structural analogue of that system and not by mentally reasoning through a set of logically related propositions.

In the following section, we will focus on a different account on thought experiments, which has been recently proposed by El Skaf (2021). This account explicitly addresses the two questions mentioned above and, as we will argue in section 4, provides a useful framework for identifying and understanding some important epistemic functions of TEs in BHT.

3.2 TEs that Reveal and Resolve Inconsistencies

In his account on TEs, El Skaf (2021) argues, contra Norton, that TEs should not be identified with arguments, even though they may contain important pieces of argumentation. In addition, he contends that the imagined experimental arrangements are not eliminable. To the contrary, they are crucial for the epistemic functions of TEs. Contra Brown, he argues that the constructive conclusion of a TE should not be understood as an inference of new *a priori* laws, but rather as a *resolution* of an inconsistency revealed by a TE, which has conjectural character. Contrary to mental model accounts, he remains pluralist as to the cognitive processes called upon when reasoning through a TE. For him, the cognitive processes can be propositional and non-propositional (El Skaf 2021, pp 6133-6135). However, the most important aspect of El Skaf's approach is that the principal functions of an important class of TEs are to "reveal" and "resolve" inconsistencies. Although most accounts of TEs in the literature would agree that some TEs reveal and resolve inconsistencies, El Skaf's account is centered around these functions and it offers a systematic analysis of the type of inconsistency revealed by a TE and the conjectural character of its possible resolutions.[8] More precisely, El Skaf (2021) identifies the following

[8] We do not exclude in this paper that some TEs could have different functions than that of revealing and resolving inconsistencies, we are merely concentrating here on TEs that do

structure in the case studies that he considers (these include Galileo's falling bodies, Maxwell's demon, Einstein's photon-boxes):

- **Step 1: Target Theoretical Question(s)** Scientists identify a target question(s) and use a TE to answer it(them).
- **Step 2: Scenario** They imagine a particular scenario, which contains a more or less well-described hypothetical or counterfactual experimental arrangement. The scenario of a TE is mainly composed of the following elements:
    1. Theoretical/empirical statements
    2. Hypothetical or counterfactual experimental arrangement, involving objects and things that happen to (or are performed by) them.
    3. Idealizations and abstractions
- **Step 3: Unfolding of the Scenario** They "unfold" the scenario, which basically means that they apply the theoretical statements involved in the experimental set-up to describe and trace the execution of the experimental arrangement.
- **Step 4: Output of the Unfolding (OU)** If the unfolding of the scenario is correctly done, they obtain a proposition as an *output.* [9]
- **Step 5: Inconsistency revealed** The interpretation of the OU can reveal a real or apparent (external) inconsistency.
- **Step 6: Inconsistency Resolved** The scientist offers a way out of the inconsistency revealed in Step 5 in the form of a *conjecture*, which is a hypothesis to be further explored and tested by future theoretical developments and, ideally, empirical confirmation.

The details of this structure are not important for our purposes, but there are three aspects of this account that will be crucial for our analysis of the epistemic role of TEs in BHT, which will be carried out in Section 4.

First, this account requires us to explicitly identify the theoretical statements that are grouped together in a TE (step 2) and to analyze their role in describing the execution of an imagined experimental arrangement (step 3). Indeed, it is mainly the application of different theoretical statements (step 3) what provides us with a result (OU) in the imagined TE. More precisely, given that it is a thought, and not an empirical, experiment, the OU is mainly obtained by applying different theoretical statements to a given experimental set-up and following their consequences through. Importantly, according to El Skaf (2021), the experimental set-up is not *eliminable*, contrary to what Norton suggests.

reveal and resolve inconsistencies.

[9] It is important to distinguish the result of such unfolding, the OU, from the conclusions of the TE (i.e. steps 5 and 6 respectively). This is sometimes conflated in the literature on TEs.

The second aspect is also related to the non-eliminability of the experimental details. Following Krimsky (1973), El Skaf (2021) distinguishes between *internal* and *external inconsistencies.* He, then, argues that the main aim of TEs is to reveal external ones. Briefly, the difference between these two kinds of inconsistencies is the following. A set of theoretical statements is said to be "internally inconsistent", if we can derive a contradiction by simply grouping these generally formulated statements together, without the need to apply them to a particular set-up. For instance, it could be argued that we get an internal inconsistency if we group together generally formulated theoretical statements from Newtonian mechanics, such as those allowing for instantaneous action at a distance, and theoretical statements from relativity theory, such as those allowing only for local action. On the other hand, a set of theoretical statements is said to be "externally inconsistent" when they do not contradict each other directly or at least in appearance, but a contradiction is manifested when they are applied to a particular set-up. For instance, when we group together Einstein's locality and separability principles with statements coming from quantum mechanics, no contradiction seems to follow. However, if these statements are confronted in a scenario such as Einstein's imagined experimental set-up (e.g. EPR and proto-EPR photon-box), an inconsistency between locality/separability and the completeness of quantum mechanics will be revealed.[10] One of the main functions, then, of the scenario of a TE is to provide an adequate hypothetical or counterfactual experimental set-up in which different theoretical statements, coming sometimes from disparate theories, can be grouped together and be confronted. We will argue in Section 4 that the case of BHT makes particularly salient that the role of many TEs is to reveal *external inconsistencies.* This is so, as we will argue, because BHT essentially groups and confronts statements coming from different theories, such as general relativity, quantum mechanics and thermodynamics, which were initially used to describe different domains and different length scales.

Finally, and more importantly, this structure clearly distinguishes between two main conclusions that can be obtained by means of a TE, that is, the revelation and the resolution of an inconsistency (steps 5 and 6). According to El Skaf (2021), each of these conclusions has its own epistemic force and merits. He points out that while the revelation of an inconsistency is "the most robust conclusion", because it clearly indicates that there is something in our theoretical web of beliefs that *must* be changed, the resolution of an inconsistency has *conjectural* character and it is best interpreted as guiding future research programs.

In the next section, we will see that the distinction between these two epistemic roles of TEs is particularly important in the case of BHT. In fact, we will go beyond El Skaf's (2021) approach by arguing that also the reasons to trust the revelation and the resolution of an inconsistency substantially differ. More precisely, we will argue that black hole TEs provide us with a hypothetical or counterfactual situation in which the domain of disparate theories that

[10] See Bokulich 2001 and El Skaf 2021 for a philosophical discussion around these TEs.

normally describe different length scales can be unified. The revelation of an inconsistency in such a scenario should be, then, taken as *conclusive knowledge*, provided that the TE is "successful"[11]. Moreover, since we are arguing that the revelation of an inconsistency should be interpreted as conclusive knowledge, this means that performing a direct empirical experiment with a similar set-up would not necessarily improve our knowledge of the alleged inconsistency between theoretical statements. In contrast, we will argue that the resolution of an inconsistency should be interpreted as *conjectural* and not as conclusive knowledge. El Skaf (2021) suggests that in order to provide evidence for a certain resolution, one should go beyond the TE. However, he does not suggest any potential ways of providing evidence for a certain resolution in cases in which direct empirical evidence is absent. Focusing on the case of BHT, as we will see next, will encourage us to consider alternative theoretical and non-empirical ways of providing evidence in such cases. In particular, we will suggest that *robustness tests*, *theoretical arguments* (such as direct calculations), and even *analogue experiments* could be potentially used to provide evidence in favor of a particular resolution in cases in which direct empirical evidence is lacking.

4 Thought Experiments in Black Hole Thermodynamics

We have seen that the philosophical debate on the epistemology of TEs has mainly focused on examples from the history of science instead of examples from ongoing physics. In this Section, we aim to expand this literature by analyzing the epistemic roles of the two TEs introduced in Section 2, as well as other TEs in BHT. We will conclude that the most important roles of black hole TEs are to unveil inconsistencies between different theoretical statements and to suggest possible ways of resolving them. We will, then, stress the conjectural character of the possible resolutions and discuss different empirical and non-empirical tools that can potentially provide evidence in favor of the plausibility of a given resolution.

4.1 Reinterpreting Wheeler's Thought Experiment

In Section 3, we explained that according to El Skaf (2021), the most important functions of many TEs are to reveal and resolve inconsistencies. A careful examination of Wheeler's TE shows that its main functions are precisely those. In fact, as we noted in Section 2, Wheeler's TE reveals an inconsistency between (i) the no-hair theorem of classical general relativity and (ii) the second law of thermodynamics. This inconsistency, and the assumption that the no-hair theorem is true, motivated Bekenstein to propose a resolution. This resolution consisted of modifying the second law of thermodynamics and introducing a

[11] There are different ways in which a TE may not be successful, for instance, if the theoretical statements are not correctly applied or the idealisations are not justified. We will come back to this in the discussion of Wheeler and Geroch's TEs carried out in Section 4.

generalized second law, which attributed an entropy to black holes that was proportional to the surface area (eq. 2). As Raphael Bousso nicely puts it: "the no-hair theorem poses a paradox, to which the area theorem suggests a resolution." (Bousso 2002, p. 830). Moreover, the following analysis of this TE shows that it nicely satisfies the structure associated to an important class of TEs, which we described in section 3.2:

- **Step 1: Target Theoretical Question(s)** Wheeler wanted to test the compatibility of thermodynamics and general relativity in the context of black holes.
- **Step 2: Scenario** He considered a counterfactual situation, recreated later by Bekenstein in 1972, in which two cups of tea were thrown into a black hole (Figure 1). This includes several auxiliary assumptions and idealizations, such as the stationarity of the black hole.
- **Step 3: Unfolding of the Scenario** Wheeler and Bekenstein later "unfolded" the scenario, which means that they applied certain theoretical statements, such as the no-hair theorem and the second law of thermodynamics to the set-up described in Figure 1.
- **Step 4: The OU** They obtained the following outcome: the total entropy of the universe may have decreased in the process.
- **Step 5: Inconsistency revealed** They interpreted the OU as a "transcendence" of the second law of thermodynamics.
- **Step 6: Inconsistency resolved** The second law was modified (generalized), so as to include the entropy of black holes. More precisely, it was reformulated as "common entropy plus black-hole entropy never decreases." (Bekenstein 1972)

Following Krimsky (1973)'s distinction between internal and external inconsistencies that we explained in Section 3, the character of the inconsistency revealed between the no-hair theorem and the second law should be rather interpreted as *external.* In fact, nothing at first sight seems to link the general relativity's no-hair theorem and the second law of thermodynamics. The first is a statement about the degrees of freedom of a black hole, whereas the second is a statement about the change in the entropy of an isolated system left to spontaneous evolution. It was rather Wheeler's TE what provided us with a counterfactual scenario, in which it was possible to confront these statements from general relativity and thermodynamics. In other words, Wheeler's scenario helped us unify the domains of these different theoretical statements, so that we could test their mutual consistency. Once this scenario was constructed, a logical inconsistency was *conclusively* revealed.

It is important to point out, however, that we are assuming here that the scenario was adequately constructed and appropriately unfolded. In fact, it is possible in principle to "block" the inconsistency revealed by a TE, for example, by showing that the theoretical statements are not adequately applied or that the idealizations are not justified. In this case, for instance, it is assumed that black holes are stationary, which means that they are “in equilibrium”. This is an idealization, which is required to formulate the laws of BHT and to characterize black holes in terms of a small number of parameters. One may question the legitimacy of this idealization, but there are some reasons to think that this idealization may be appropriately justified. (Heusler 1996, Wald 2001)

According to Bekenstein (1972), the apparent inconsistency between the no-hair theorem and the second law cried for a resolution. He says: entropy is “necessitated by [Wheeler’s TE]. Without it the second law is definitely transcended. With black-hole entropy the second law becomes a well-defined statement susceptible to verification by an exterior observer.” (Bekenstein 1972, p. 738) As we mentioned in Section 2.1, Bekenstein suggested, then, to generalize the second law by attributing entropy to black holes, a quantity that was proportional to its area. However, we should note that nothing in the TE, or in any TE for that matters, forces us to accept a specific resolution. In this particular case, nothing in Wheeler’s TE logically forces us to modify or generalize the second law, so as to include the entropy of the black hole. In fact, Bekenstein’s proposal initially appeared to be largely speculative and physically implausible.[12] Indeed, the attribution of an entropy to black holes was for many physicists counterintuitive, since it appeared to relate a mathematical theorem in differential geometry, namely the area theorem, with a statistical law, namely the second law of thermodynamics (Wald 2001). Furthermore, for some scientists, the apparent tension between the no-hair theorem of general relativity and the second law of thermodynamics did not even required a resolution. Wald (2019), for instance, says:

> My own view at the time was that the second law of thermodynamics is a statistical law, not a fundamental law, so its “transcendence” would be more palatable than the transcendence of an apparently fundamental law like baryon conservation. Thus, I was quite comfortable with the transcendence of the second law of thermodynamics. But Wheeler did not feel this way. (Wald 2020, p. 5)

The above shows that other resolutions for Wheeler’s TE were possible in principle. The most straightforward one would have been simply to bit the bullet and accept that the second law was transcended, which appears to correspond to Wald’s initial attitude towards this problem at the beginning of the 1970’s. Another reason to be suspicious about Bekenstein’s resolution was the belief that black holes were systems at absolute zero temperature, which was also

[12] Almeida (2021), for instance, suggests that Bekenstein’s resolution of Wheeler’s TE may have been inspired by Brillouin’s resolution of Maxwell’s demon, a well known series of TEs in physics. More precisely, in 1950, Leon Brillouin proposed a resolution of Maxwell’s TE based on information theory. According to Almeida, this inspired Bekenstein to address Wheeler’s TE, which Bekenstein named “Wheeler’s demon”, in a similar way. (See Earman and Norton 1998,1999, Norton 2005, 2013, El Skaf 2017, for a philosophical analysis of Maxwellian demons TEs).

supposed to be a consequence of Geroch's TE (see Section 4.2). As Wald (2020) puts it:

> However, at the time, I felt that this was an utterly ridiculous project to work on. First, as already mentioned, I was not troubled by the apparent transcendence of the second law of thermodynamics. Second, the analogy between the second law and the area theorem seemed extremely artificial; it seemed quite unnatural to me to try to marry a statistical law with a mathematical theorem. But, most importantly, in the absence of a fully developed quantum theory of gravity, what could one possibly show and/or how could one possibly argue for the validity of any highly speculative ideas on black hole entropy that one might propose? (p. 5)

An important question is, then, what convinced Wald and an important part of the scientific community working in BHT that Bekenstein's resolution was the correct solution of Wheeler's TE. A careful examination of the discussion around possible resolutions of Wheeler's TE shows that it was principally Hawking's prediction that black holes emit radiation with temperature proportional to the surface gravity what convinced them about the plausibility of Bekenstein's resolution. Wald (2020) makes this explicit, when he says:

> Then, a miracle occurred! In 1974, Hawking calculated particle creation effects for a body that collapses to a black hole, and he made the amazing discovery that a distant observer will see a steady, thermal distribution of particles emerge at a temperature $T = k/2\pi$. So, a black hole truly has a nonzero physical temperature proportional to its surface gravity! Black hole thermodynamics now appeared to be entirely consistent. In particular, if one placed a black hole in a radiation bath of temperature $T_{bath}k/2\pi$ the black hole radiation would dominate over absorption, and the generalized second law would hold. The entropy $S_{BH} = A/4$ could now be interpreted as the physical entropy of the black hole – with the unknown constant in Bekenstein's original proposal now fixed by the value of the Hawking temperature. Bekenstein was right! (p. 7)

It is important to point out, however, that Hawking's prediction was not the result of an empirical observation, but rather of what appeared to be an unimpeachable mathematical derivation. The analysis of Geroch's TE in the next sections will reveal that apart from mathematical derivations, robustness tests (the possibility to replicate the experiment under different conditions) and analogue experiments may also play a role in the acceptance of a given resolution.

4.2 Understanding the Role of Geroch's Engine in BHT

Let us now return to Geroch's engine TE. In this TE, we are grouping together statements from classical general relativity, thermodynamics, and quantum mechanics. We explained that, classically, if we lower the box close to the horizon before dropping the radiation in, one could recover all of the energy that was originally in the box as "work" (Figure 2). Since no energy would be delivered to the black hole, the first law of black hole thermodynamics implies that the black hole area A would not increase. However, this is in contradiction with Bekenstein's formulation of GSL, which associates the entropy of a black hole with the area (eq. 2). Furthermore, since in Geroch's TE, it is possible in principle to convert heat into work with 100% efficiency, this implies that the physical temperature of a black hole is absolute zero. This is in contradiction with the assignment of finite non-zero temperature to the black hole, as required by the first law of black hole thermodynamics, if one assigns a finite non-zero entropy to the black hole (details in Wald 2001). In sum, it seems that the main function of Geroch's TE is to reveal a contradiction between the *properties of the horizon according to classical general relativity* (such zero energy and zero temperature) and both the *first* and the *generalized second law of black hole thermodynamics* as well as to suggest possible ways of resolving them. Furthermore, like Wheeler's TE, we see that Geroch's TE also has the structure associated with TEs that we described in Section 3.2:

- **Step 1: Target Theoretical Question(s)** Geroch probably wanted to test the compatibility between thermodynamics and general relativity in the context of black holes.

- **Step 2: Scenario** He considered a counterfactual situation, in which one lowers a box filled with radiation of high entropy matter all the way to the horizon of the black hole before dropping the radiation in (Figure 2). This situation involves several auxiliary assumptions and idealizations, such as the stationarity of the black hole and the stationarity of the entire spacetime.

- **Step 3: Unfolding of the Scenario** Geroch "unfolded" the scenario, which means that he applied certain theoretical statements, such as the properties of the horizon according to classical general relativity and the laws of black hole thermodynamics.

- **Step 4: The OU** He concluded that in that particular scenario, all of the "heat" of the matter could be converted to "work" in the laboratory from which one did the lowering.

- **Step 5: Inconsistency revealed** Bekenstein (1972) interpreted the OU as if, in this particular scenario, it was possible to run a Carnot cycle with 100% efficiency, which was in contradiction with the second law of thermodynamics. Furthermore, since no energy would be delivered to the black hole, and the first law of black hole thermodynamics implies that the black hole area A would not increase, he later (1973) interpreted this result as contradicting his GSL, which associates the entropy of a black hole with the area.

- **Step 6: Inconsistency resolved** Bekenstein (1981) imposes a physical bound (nowadays known as "Bekenstein bound") that cannot be exceeded by the box or any other physical system. Other alternative resolutions have also been suggested, most notably by Unruh and Wald (1982, 1983).

As in the case of Wheeler's TE, the scenario helps us conceive a hypothetical situation in which we unify the domain of theories that were usually used to describe disparate domains, such as thermodynamics, classical general relativity and also quantum mechanics. Applying theoretical statements to this scenario (unfolding the scenario) reveals a logical contradiction between statements of general relativity and black hole thermodynamics, which was well hidden before running the TE. Like the case of Wheeler's TE, if we assume that the experiment is successful, that is, if we assume that the theoretical statements are adequately applied and the idealizations are justified, we should interpret the revelation of this inconsistency as *conclusive knowledge.* It is important to point out, however, that it is always possible to "block" the inconsistency by challenging some of the assumptions made in the scenario. For instance, an assumption that has been matter of controversy is the *stationarity of the entire spacetime.* The assumption of stationarity of the black hole and stationarity of the entire spacetime appears to be essential to relate changes in quantities defined at the horizon (like the area) to changes of quantities defined at infinity (like the mass and the angular momentum). However, one would expect that the equilibrium behavior of a black hole would require only a form of local stationarity at the horizon, which would allow one to formulate the first law of black hole thermodynamics in terms of local definitions of quantities like mass and angular momentum at the horizon (Wald 2001). The latter motivated Lewandowski (2000) to replace the stationarity assumption by the notion of an *isolated horizon*, which does not require the entire spacetime to be stationary.

In contrast to the revelation of an inconsistency by means of Geroch's TE, resolving this inconsistency has an intrinsic conjectural character. Indeed, the fact that there is still no consensus with respect to the best resolution of Geroch's TE makes the speculative character of potential resolutions particularly clear. In contrast to Wheeler's TE, there has been no resolution of Geroch's TE that has been widely accepted by the scientific community working on black holes. We believe that one of the reasons for this is that mathematical derivations do not favour one resolution over the others. For instance, Hawking's derivation supported Bekenstein's insight that the temperature associated to black holes corresponded to a truly a physical temperature, but it did not solve the paradox revealed by Geroch's TE. More to the point, if one could lower the box arbitrarily close to the horizon, one could still get rid of the entropy without increasing the area. In this case, Hawking's prediction that black holes radiate would not help, since for arbitrarily large black holes, quantum effects and the Hawking temperature would be arbitrarily small (Wald 2020).

Since mathematical derivations have not offered a compelling and rebuttal argument supporting a particular resolution of Geroch's TE yet, robustness arguments have played a central role in the discussion. We take robustness tests as repetitions of the TE under slightly different circumstances, for instance, changing the shape of the box in Geroch's TE. Our analysis in Section 2 shows that robustness tests have been used from both sides, that is, from Bekenstein's side as well as Unruh and Wald's side to invalidate alternative resolutions. As we explained in Section 2.3, Wald and Unruh (1982, 1983) presented different counterexamples of Bekenstein bound that would challenge the robustness and generality of Bekenstein's resolution. As Page (2005) puts it:

> Perhaps the main difficulty is how to give precise definitions for the system and for its S, E and R (Bekenstein 1982). For various choices of those definitions, one could easily come up with counterexamples to the conjecture. (Page 2005, p. 12)

Similarly, Bekenstein (1983, 1994, 1999) tried to invalidate Unruh and Wald's (1982) results by pointing out situations in which the appeal to a bouyancy force would not suffice to prevent a violation of GSL. Although counterarguments have been given, the discussion remains open until now (Wald 2001, 2020, Pelath and Wald 1999).

Apart from robustness tests, another tool that could possible help supporting a particular resolution when direct empirical experiments are not available is the use of analogical reasoning and, in particular, so called "analogue experiments". We will briefly discuss the role of analogue experiments in BHT in Section 4.4.

4.3 Other TEs in Black Hole Physics: From the Information Loss Paradox to Firewalls

So far, and for simplicity, we have focused on two TEs in BHT. There are, however, several TEs that have occupied an important role in BHT. Perhaps the most important one is Hawking TE. This TE has been sometimes referred to as the "mother of all thought experiments, one that still keeps physicists awake at night" and "perhaps Dr. Hawking's most profound gift to physics" (Carroll, 2018).

We do not have the space to analyze this TE in detail here, but a short analysis will suggest that it fulfils the same epistemic roles as Wheeler and Geroch TEs. Put simply, Hawking TE consists of imagining throwing a bit of information, such as a book, a computer, even an elementary particle, into a black hole.[13] We then ask what would happen to the information contained in the thrown object, especially after the complete evaporation of the black hole, which according to Hawking (1974) would occur in a finite time. The answer to this question leads to a paradox, called in the literature "The Information Loss Paradox".

The Information Loss Paradox can be described in more detail as follows. On the one hand, according to classical general relativity, the matter responsible for the formation of a black hole propagates into a singularity lying within

[13] This TE was implicitly formulated in Hawking 1976, but reconstructed in this way by Susskind 2008, among others.

the deep interior of the black hole, where gravity is so intense that nothing can escape it. On the other hand, the semiclassical framework – which is a hybrid approach used by Hawking that considers quantum field theory on curved spacetime – predicted that quantum correlations between the exterior and the interior continuously build up as the black hole evaporates. In fact, these correlations played a crucial role in the derivation of Hawking radiation. Since the matter that falls into a black hole could possess quantum correlations with matter that remains outside of the black hole, it is difficult to conceive how these correlations could be restored during the process of black hole evaporation in a way that is consistent with general relativity. So, either there is a mechanism that restores the correlations during the late stages of the evaporation process, which may contradict one of the main principles of general relativity, or, by the time the black hole has evaporated completely, an initial pure state would have evolved into a mixed state, that is, information would have been lost. The latter is commonly said to be in conflict with quantum mechanics.[14] Hawking (1976) concluded from this TE that information has been lost, thus challenging some of the fundamental principles of quantum mechanics:

> The conclusion of this paper is that gravitation introduces a new level of uncertainty or randomness into physics over and above the uncertainty usually associated with quantum mechanics. Einstein was very unhappy about the unpredictability of quantum mechanics because he felt that "God does not play dice." However, the results given here indicate that "God not only plays dice, He sometimes throws the dice where they cannot be seen." (Hawking 1976, p. 13-14)

However, Hawking's conclusion was not the only possible resolution of the Information Loss Paradox. Different resolutions have been proposed such as Maldacena's "AdS/CFT" duality (Maldacena 1999), the holographic principle ('t Hooft 1988, Susskind 1995), and Susskind complementarity (Susskind and Thorlacius, 1993; Stephens et al. 1994). Some of these resolutions seem to favour general relativity, whereas others are more conservative towards quantum mechanics. As Susskind (2008) nicely puts it in his popular book *The Black Hole War*:

> The Black Hole War was a genuine scientific controversy [...] Eminent theoretical physicists could not agree on which principles of physics to trust and which to give up. Should they follow Hawking, with his conservative views of space-time, or 't Hooft and myself, with our conservative views of Quantum Mechanics? Every point of view seemed to lead only to paradox and contradiction. (Susskind 2008, p. 9)

[14] The are two reasons why it is claimed that the evolution of a pure state into a mixed state is in conflict with quantum mechanics: i) Such evolution is incompatible with a fundamental principle of quantum theory, which postulates a unitary time evolution of a state vector in a Hilbert space, and ii) such evolution give rise to violations of causality/conservation of energy/momentum (see Wald 2001 for details).

More recently, Almheiri et al (2013) re-interpreted the information loss paradox as a contradiction between the following three statements: (i) Hawking radiation is in a pure state, (ii) the information is lost, and (iii) the infalling observer does not encounter anything unusual at the horizon. They point out that the most conservative resolution is to give up (iii) and conclude that the infalling observer finds a "Firewall", which means that it burns up at the horizon. Although this resolution implies some elements of nonlocality, they show that other alternatives may cause notable violations of the semiclassical framework. More recent resolutions include extensions of Maldacena's resolution appealing to "AdS/CFT" duality (Almheiri 2018).

Interestingly, Joseph Polschinski, one of the proponents of the Firewalls resolution, in a talk entitled "Black Holes, Quantum Mechanics and Firewalls" delivered in November 2013 at a Simons Symposium, explicitly emphasized the role of TEs as useful tools for exposing incompleteness and inconsistencies and analysed the information loss paradox as such.[15] He says:

> The theories of quantum mechanics and general relativity are each very well tested and successful in their own regimes. Thought experiments can expose inconsistencies. Black holes have proven to be useful arenas for the confrontation quantum mechanics and general relativity.

In sum, it appears that the principal role of Hawking TE and other TEs taking place in the discussion around the Information Loss Paradox was precisely to unveil a paradox between crucial statements of quantum mechanics and general relativity, which was well hidden behind the theories. As Susskind (2008) puts it:

> Theoretical physicists are struggling to gain a foothold in a strange land. As in the past, thought experiments have brought to light paradoxes and conflicts between fundamental principles. This book is about an intellectual battle [with Hawking on the resolution of the Information Loss Paradox] over a single thought experiment. (Susskind 2008, p. 8)

Furthermore, a brief examination of the discussion that follows Hawking's insights shows the conjectural character of the debate, which remains open until now. This is in tune with our analysis on Wheeler and Geroch's TEs.

4.4 On the Use of Direct Calculations and Analogue Experiments to Test Resolutions

We have previously argued that TEs can potentially lead to conclusive knowledge with respect to the inconsistency between different theoretical statements. However, we have also stressed that they are incapable of helping us test the

[15] Talk available at https://www.youtube.com/watch?v=424rxT`bVlwt=635s

validity of a proposed resolution to a given inconsistency. In fact, as we saw it in the case of Wheeler's demon, the TE was capable to convince the scientific community that there was an inconsistency between statements of general relativity and thermodynamics, when applied to a particular scenario, but it was incapable of convincing the entire scientific community about a particular resolution for this inconsistency. In fact, what finally convinced many physicists about Bekenstein's attribution of entropy to black holes was mainly the consistency of this resolution with Hawking's derivation. The latter suggests that in absence of direct empirical evidence, mathematical derivations may play important role in accepting particular resolutions, and, in this sense, can be complementary to the use of TEs. For instance, Wallace (2018, 2019) argues, in line with many physicists (e.g. Belgiorno et al. 2019), that the confidence of the scientific community in BHT rests principally on independent calculations performed with different premises and different approximations, which led to the same results.

Another promising candidate for giving support to a particular resolution in the absence of direct empirical evidence are analogue experiments, which are material experiments performed not on the target system, but on a source analogous system. In the past forty years, physicists have tried to give *some* empirical support to the predictions of BHT, by reproducing the characteristics of an event horizon in an analogous physical system, which is simple enough to be run in the laboratory, such as condensed matter systems. This new research program has been called "analogue gravity"[16]. There has been an important discussion in the philosophy of science as to whether analogue experiments performed in the context of analogue gravity can have confirmatory power and, in particular, if they can provide us with evidence of the same kind as direct experiments (e.g. Dardashti et al. 2017, 2019; Thébault 2019; Crowther et al. 2021). This is not the place to review this discussion in detail, but if one agrees that analogue experiments can provide at least *some* empirical support for the predictions of BHT, then one may also believe that they can play a role in the acceptance of a given resolution. For instance, the argument that appeared to have convinced both Unruh and Wald that a bouyancy force may arise close to the horizon was that it became apparent that there would be a real buoyancy effect associated to Unruh radiation, which was taken as an analog of Hawking radiation. Page (2005) says:

> [H]e [Wald] and Unruh independently rediscovered this mechanism [bouyancy force] after realizing that the Unruh acceleration radiation would make the buoyancy effect real. (Page 2005, p. 13)

Recently, real analogue experiments have been performed to test Unruh effect with classical analogues (Blencowe and Wang 2020, Leonhardt et al. 2018).[17] Additionally, many analogue experiments have relied on Maldacena's

[16] For a review of the literature on analogue gravity see Barceló et al. 2005; Faccio et al. 2013; Belgiorno et al. 2019.

[17] See Gryb et al. (2021) for a discussion on the problems that may arise when we associate Unruh effect with Hawking effect.

AdS/CFT duality (Bilić, et al. 2015; Dey, R. et al. 2016) and there have been some attempts to test the Firewall resolution by considering fluid analogues (Pontiggi 2015).

The extent to which analogue experiments can actually provide us with genuine evidence for particular resolutions in BHT needs to be further investigated and should probably be evaluated on a case-by-case basis. However, it suffices for our purposes to have shown that the conjectural character of candidate resolutions invites us to consider alternative non-empirical or surrogative means, especially in cases in which the phenomenon under investigation is beyond the reach of direct experimentation, such as the case of black holes.

5 Conclusion

Polchinski concludes his 2013 lecture by saying:

> Thought experiments with black holes have led to some surprising discoveries: black hole bits, the holographic principle, Maldacena's duality. The latest thought experiment presents new challenges, and we can hope that it will lead us to a more complete theory of quantum gravity. (Polchinski, 2013)

We share Polchinski's enthusiasm and we agree that the importance of TEs in BHT should be acknowledged. However, we also believe that it is important to specify the power and limits of TEs. We have argued throughout this paper that the principal functions of TEs in BHT, like many other TEs from the history of physics, is to reveal and resolve external inconsistencies. We stressed that whereas the revelation of an inconsistency provides conclusive knowledge, the resolution is only conjectural.

When one focuses on historical case studies, it is very difficult to see the conjectural character of a given resolution, especially if the alternative resolutions did not last long or were not pursued for a reason or another. In contrast, analyzing the use of TEs in ongoing physics allows one to see, before the end of the inquiry, the highly conjectural character of different resolutions. In addition, we have seen that black holes are an ideal arena to understand the importance of different empirical and non-empirical tools in the absence of direct empirical evidence. In particular, we have stressed that *robustness tests*, *theoretical arguments* (such as direct calculations) and *analogue experiments* may play a role in the acceptance or rejection of a given resolution, and so complement the knowledge obtained on the basis of TEs.

Acknowledgments

We are grateful to Giovanni Valente and Christian Wüthrich for fruitful discussions around the main ideas of this article. We would also like to thank Robert Wald and Juan Maldacena for inspiring discussions on the importance of thought experiments in black hole physics at the Seven Pines Symposium 2022. We are also grateful to the audiences of the 20th European Conference on Foundations of Physics 2021 and the BSPS conference 2022. Finally, we would like to thank two anonymous referees for their helpful comments.

References

Almeida, C.R. (2021). The thermodynamics of black holes: from Penrose process to Hawking radiation. *Eur. Phys. J. H* 46:20.

Almheiri, A. (2018). Holographic Quantum Error Correction and the Projected Black Hole Interior. arXiv:1810.02055

Almheiri, A., Marolf, D., Polchinski, J. and Sully J. (2013) Black holes: complementarity or firewalls?. *J. High Energ. Phys.* 2013, 62.

Barceló et al. (2005). Analogue Gravity. *Living Rev. Relativ.* 8, 12

Bekenstein, J.D. (1972). Black Holes and the Second Law. *Lettere al Nuovo Cimento*, 4(15): 737–740.

Bekenstein, J.D. (1973). Black Holes and Entropy. *Physical Review D*, 7, 2333-2346.

Bekenstein, J.D. (1974). Generalized second law of thermodynamics in black-hole physics. *Physical Review D*, 9, 3282-3300.

Bekenstein, J.D. (1980). Black-Hole Thermodynamics. *Physics Today*, 33, 24-31.

Bekenstein, J.D. (1981). Universal Upper Bound on the Entropy-to-Energy Ratio for Bounded Systems. *Physical Review D*, 23, 287–298.

Bekenstein, J.D. (1983). Entropy bounds and the Second Law for Black Holes. *Physical Review D*, 27, 2262-2270

Bekenstein, J.D. (1984). Entropy Content and Information Flow in Systems with Limited Energy. *Physical Review D*, 30, 1669-1679

Bekenstein, J. D. (1994). Do we understand black hole entropy?. arXiv preprint gr-qc/9409015.

Bekenstein, J.D. (1999). Non-Archimedian Character of Quantum Buoyancy and the Generalized Second Law of Thermodynamics. *Physical Review D*, 60, 124010.

Belgiorno et al. (2010). Hawking Radiation from Ultrashort Laser Pulse Filaments. *Phys. Rev. Lett.* 105, 203901

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

Blencowe Miles P. and Wang H. (2020) Analogue gravity on a superconducting chip. *Phil. Trans. R. Soc. A*.37820190224

Bilić, et al (2015). Analog geometry in an expanding fluid from AdS/CFT perspective. *Physics Letters B*, 743 pp. 340-346

Bokulich, A. (2001). Rethinking thought experiments. *Perspectives on Science*, 9(3), 285–307.

Bousso, R. (2002). The holographic principle. *Rev. Mod. Phys.* 74, 825.

Brendel, E. (2018). The argument view: are thought experiments mere picturesque arguments? In Stuart, M.T. et al. (Eds.), *The Routledge Companion to Thought Experiments* (pp. 281–293). London and New York: Routledge.

Brown, J.R. (1991). *Laboratory of the mind: thought experiments in the natural sciences*, 1st edition. London: Routledge.

Carroll, S. (2018). Stephen Hawking's Most Profound Gift to Physics. New York Times, https://www.nytimes.com/2018/03/15/opinion/stephen-hawking-quantum-gravity.html

Christodoulou, D. (1970). Reversible and Irreversible Transformations in Black-Hole Physics. *Physics Review D* 94, 104068.

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

Curiel, E. (2014). Classical black holes are hot Preprint at https://arxiv.org/abs/1408.3691

Dardashti, R., Thébault K., and Winsberg E. (2017). Confirmation via Analogue Simulation: What Dumb Holes Could Tell Us about Gravity. *British Journal for the Philosophy of Science*, 68, 55–89.

Dardashti, R., Hartmann, S., Thébault K., and Winsberg E. (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.

Dey, R. et al. (2016). AdS and dS black hole solutions in analogue gravity: The relativistic and non-relativistic cases. *Physical Review D* 94, 104068.

Dougherty, J., and Callender, C. (forthcoming). Black hole thermodynamics: More than an analogy? In A. Ijjas and B. Loewer (Eds.), *Guide to the Philosophy of Cosmology.* Oxford: Oxford University Press.

Earman, J. (2011). The Unruh effect for philosophers. *Studies in History and Philosophy of Science Part B: Studies in History and Philosophy of Modern Physics*, 42 (2):81-97.

Earman, J. and Norton, J. (1998). Exorcist XIV: The wrath of Maxwell's demon. Part I. From Maxwell to Szilard. *Studies in the History and Philosophy of Modern Physics.* 29(4), 435–471.

Earman, J. and Norton, J. (1999). Exorcist XIV: The wrath of Maxwell's demon. Part II. From Szilard to Landauer. *Studies in the History and Philosophy of Modern Physics*, 30(1).

El Skaf, R. (2021). Probing theoretical statements with thought experiments. *Synthese* 199, 6119–6147.

El Skaf, R. (2017). What notion of possibility should we use in assessing scientific thought experiments? *Lato Sensu: Revue De La Société De Philosophie Des Sciences* , 4(1), 19–30.

El Skaf, R. (2018). The function and limit of Galileo's falling bodies thought experiment: Absolute weight, specific weight and the medium's resistance. *Croatian Journal of Philosophy*, XVII(52), 37–58.

Evans, P. W., and Thébault, K. P. (2020). On the limits of experimental knowledge. *Philosophical Transactions of the Royal Society A*, 378(2177), 20190235.

Faccio D. et al. (2013). *Analogue gravity phenomenology.* Berlin, Germany: Springer International Publishing.

Gendler, T. S. (1998). Galileo and the indispensability of scientific thought experiment. *The British Journal for the Philosophy of Science*, 49, 397–424.

Gryb. S, Palacios. P and Thébault, K. (2021) On the Universality of Hawking Radiation. *The British Journal for the Philosophy of Science*, 72(3), 809-837.

Heusler, M. (1996) *Black Hole Uniqueness Theorems.* Cambridge University Press, Cambridge.

Hawking, S. (1971). Gravitational Radiation from Colliding Black Holes.*Physical Review D* 94, 104068.

Hawking, S. (1974). Black Hole Explosions?. Nature 248: 30-31.

Hawking, S. (1976). Breakdown of predictability in gravitational collapse. *Physical Review D.* 14, 2460-2473.

Krimsky, S. (1973). The use and misuse of critical Gedankenexperimente. *Zeitschrift für allgemeine Wissenschaftstheorie*, 4, 323–334.

Leonhardt et al. (2018). Classical analog of the Unruh effect *Physical Review A* 98, 022118

Lewandowski. J. (2000). Space-times admitting isolated horizons. *Class.Quantum Grav.* 17, L53-L59.

Maldacena, J. (1999). The Large-N Limit of Superconformal Field Theories and Supergravity. *International Journal of Theoretical Physics* 38, 1113–1133

Matsas G.E.A and Rocha da Silava A.R. (2005) New thought experiment to test the generalized second law of thermodynamics. *Physical Review D*, 71, 107501.

Miščević N. (1992). Mental models and thought experiments. *International Studies in the Philosophy of Science* 6, 215-226.

Nersessian, N. J. (1993). In the theoretician's laboratory: thought experimenting as mental modelling. In D.Hull, M. Forbes and K. Okruhlik (Eds.), PSA 1992 (Vol. 2, pp. 291–301). Philosophy of Science Association: East Lansing.

Nersessian, N. J. (2007). Thought experiments as mental modelling: Empiricism without logic. *Croatian Journal of Philosophy* VII: 125–161.

Norton, J. (2005). Eaters of the lotus: Landauer's principle and the return of Maxwell's demon. *Studies in the History and Philosophy of Modern Physics*, 36(2), 375–411.

Norton, J. D. (1991). Thought experiments in Einstein's Work. In T. Horowitz G. Massey (Eds.), *TEs in Science and Philosophy* (pp. 129–148). Lanham: Rowman and Littlefield.

Norton, J. D. (1996). Are thought experiments just what you thought? C*anadian Journal of Philosophy,* 26, 333–366.

Norton, J. D. (2004). Why thought experiments do not transcend empiricism. In C. Hitchcock (Eds.), *Contemporary debates in the philosophy of science* (pp. 44–66). Oxford: Blackwell.

Palmieri, P. (2005). 'Spuntur lo scoglio più duro': did Galileo ever think the most beautiful thought experiment in the history of science? *Studies in History and Philosophy of Science* 36, 305–322.

Page, D.N. (2005). Hawking radiation and black hole thermodynamics. New J. Phys. 7 203

Page, D.N. (2020). The Bekenstein Bound. In Brink, L., Mukhanov, V.F., Rabinovici, E., Phua, K. K. (Eds.). *Jacob Bekenstein: The Conservative Revolutionary.* MA, US: World Scientific. pp. 159-171.

Pelath, M.A., and Wald, R.M. (1999). Comment on Entropy Bounds and the Generalized Second Law. *Physical Review D*, 60, 104009.

Polchinski, J. (2017). The Black Hole Information Problem. In Polchinski et al. (Eds.) *TASI 2015: New Frontiers in Fields and Strings: Proceedings of the 2015 Theoretical Advanced Study Institute in Elementary Particle Physics.* Boulder, Colorado, 1-26 June 2015 (pp. 353-397). World Scientific.

Pontiggia, L. (2015). Firewall Argument for Acoustic Black Holes. Masters thesis, Physics Department, Unversity of the Witwatersrand.

Ruffini, R., Wheeler, J.A. (1971). Introducing the black hole. *Physics Today*, 24. 30-41.

Susskind, L. (1995). The world as a hologram. Journal of Mathematical Physics, 36(11), pp. 6377-6396.

Susskind, D. and Thorlacius L. (1993). Gedanken Experiments involving Black Holes. arXiv:hep-th/9308100

Stuart, M.T. (2018). How Thought Experiments Increase Understanding. In Stuart, M.T. et al. (Eds.), *The Routledge Companion to Thought Experiments.* London: Routledge (pp. 526-544).

Susskind, L. (2008). *The Black Hole War: My Battle With Stephen Hawking to Make the World Safe for Quantum Mechanics.* New York: Little, Brown.

Thébault K. (2019). What Can We Learn from Analogue Experiments? In R. In Dardashti, R., Dawid, R., and Thébault, K. (Eds.) *Why Trust a Theory?: Epistemology of Fundamental Physics* (pp. 184-201). Cambridge: Cambridge

University Press.

Themes in Contemporary Physics II. Essays in honor of Julian Schwinger's 70th birthday, pp. 77-89.

Unruh, W.G. and Wald R.M. (1982). Acceleration Radiation and the Generalized Second Law of Thermodynamics. *Physical Review D*, 25, 942–958.

Unruh, W.G. and Wald R.M. (1983). Entropy Bounds, Acceleration Radiation and the Generalized Second Law. *Physical Review D*, 27, 2271-2276.

Wüthrich, C. (2019). Are Black Holes about Information?. In Dardashti, R., Dawid, R., and Thébault, K. (Eds.) *Why Trust a Theory?: Epistemology of Fundamental Physics* (pp. 202-223). Cambridge: Cambridge University Press.

Wald, R.M. (2001). The Thermodynamics of Black Holes. *Living Rev.* in Rel. 4, 6.

Wald, R.M. (2020). Jacob Bekenstein and the Development of Black Hole Thermodynamics. In Brink, L., Mukhanov, V.F., Rabinovici, E., Phua, K. K. (Eds.). *Jacob Bekenstein: The Conservative Revolutionary.* MA, US: World Scientific, pp. 3-10.

Wallace, D. (2018). The Case for Black Hole Thermodynamics, Part I: Phenomenological Thermodynamics. *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 Modern Physics*, 66: 103–117.

Weinstein, G. (2021). Demons in Black Hole Thermodynamics: Bekenstein and Hawking. https://arxiv.org/abs/2102.11209v2